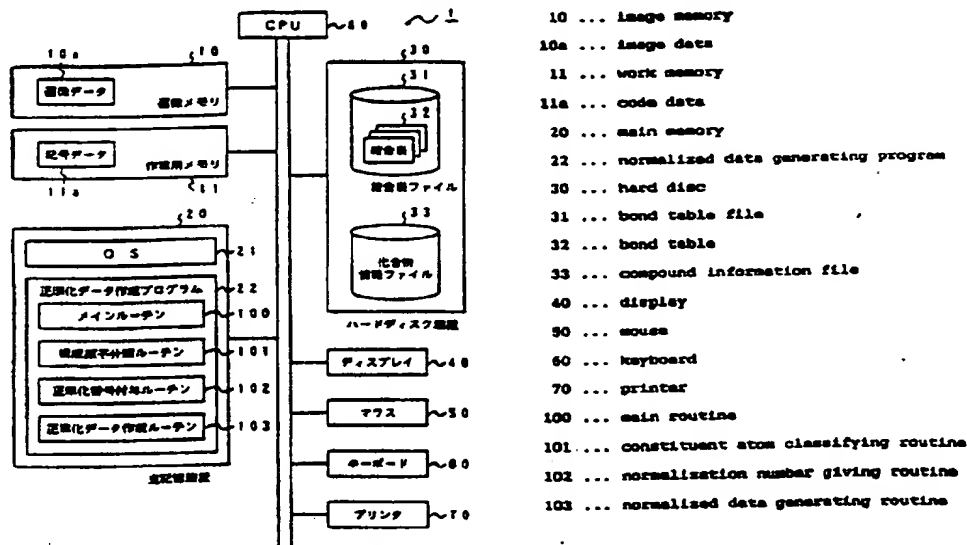




<p>(51) 国際特許分類 G06F 17/00</p>	<p>A1</p>	<p>(11) 国際公開番号 WO97/44744</p> <p>(43) 国際公開日 1997年11月27日 (27.11.97)</p>
<p>(21) 国際出願番号 PCT/JP97/01661</p> <p>(22) 国際出願日 1997年5月16日 (16.05.97)</p> <p>(30) 優先権データ 特願平8/125117 1996年5月20日 (20.05.96) JP 特願平8/125123 1996年5月20日 (20.05.96) JP</p> <p>(71) 出願人 (米国を除くすべての指定国について) 呉羽化学工業株式会社 (KUREHA KAGAKU KOGYO KABUSHIKI KAISHA)[JP/JP] 〒103 東京都中央区日本橋堀留町一丁目9番11号 Tokyo, (JP)</p> <p>(72) 発明者; および (75) 発明者/出願人 (米国についてのみ) 朝永 淳(TOMONAGA, Atsushi)[JP/JP] 〒180 東京都武蔵野市境南町三丁目17-13 Tokyo, (JP)</p> <p>(74) 代理人 弁理士 長谷川芳樹, 外(HASEGAWA, Yoshiki et al.) 〒104 東京都中央区京橋二丁目13番10号 京橋ナショナルビル6F 創英国際特許事務所 Tokyo, (JP)</p>		<p>(81) 指定国 CA, KR, US, 欧州特許 (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).</p> <p>添付公開書類 国際調査報告書</p>

(54) Title: **NORMALIZED DATA GENERATOR, NORMALIZED DATA GENERATING METHOD AND RECORDING MEDIUM FOR GENERATING NORMALIZED DATA**

(54) 発明の名称 正準化データ作成装置、正準化データ作成方法及び正準化データ作成用記憶媒体



(57) Abstract

A normalized data generator comprising an input means which receives input of characteristic data related to the atoms of which a compound is composed and data on bond pairs between atoms, and a normalized data generating means which generates normalized data by which the chemical structure of the compound can be uniquely specified in accordance with the characteristic data and the bond pair data. The normalized data generating means has a first processing unit which classifies respective equivalent atoms into different classes in accordance with the characteristic data and the bond pair data and gives the atoms class numbers different with the classes, a second processing unit which gives the atoms normalization numbers uniquely corresponding to the structure of the compound in accordance with the class numbers and a third processing unit which generates the normalized data in accordance with the normalization numbers.

(57) 要約

化合物を構成する各原子についての固有データ及び原子間の結合対データの入力を受け付ける入力手段と、前記入力手段で受け付けられた固有データ及び結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する正準化データ作成手段とを備えた正準化データ作成装置である。前記正準化データ作成手段は、前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える第1の処理部と、前記第1の処理部で各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える第2の処理部と、前記第2の処理部で各原子に与えられた正準化番号に基づいて、前記正準化データを作成する第3の処理部とを備える。

参考情報

PCTに基づいて公開される国際出願のパンフレット第一頁に記載されたPCT加盟国を同定するために使用されるコード

AL	アルバニア	ES	スペイン	LR	リベリア	SG	シンガポール
AM	アルメニア	FI	フィンランド	LS	レソト	SI	スロヴェニア
AT	オーストリア	FR	フランス	LT	リトアニア	SK	スロバキア共和国
AU	オーストラリア	GA	ガボン	LU	ルクセンブルグ	SL	シエラレオネ
AZ	アゼルバイジャン	GB	英国	LV	ラトヴィア	SN	セネガル
BA	ボスニア・ヘルツェゴビナ	GE	グルジア	MC	モナコ	SZ	スワジランド
BB	バルバドス	GH	ガーナ	MD	モルドヴァ共和国	TD	チャード
BE	ベルギー	GM	ガンビア	MG	マダガスカル	TG	トーゴ
BF	ブルキナ・ファソ	GN	ギニア	MA	マダガスカル	TJ	タジキスタン
BG	ブルガリア	GR	ギリシャ	ME	マケドニア旧ユーゴスラヴィア共和国	TM	トルクメニスタン
BJ	ベナン	GU	グアム	MK	マケドニア旧ユーゴスラヴィア共和国	TR	トルコ
BR	ブラジル	HN	ホンジュラス	ML	マリ	TT	トリニダード・トバゴ
BS	バハマ	ID	インドネシア	MN	モンゴル	UA	ウクライナ
BT	ブータン	IE	アイルランド	MR	モリタニア	UG	ウガンダ
CA	カナダ	IL	イスラエル	MW	マラウイ	US	米国
CC	中央アフリカ共和国	IS	アイスランド	MX	メキシコ	UZ	ウズベキスタン
CD	コンゴ	IT	イタリア	NE	ニジェール	VN	ヴェトナム
CE	スイス	JP	日本	NL	オランダ	YU	ユーゴスラビア
CH	スイス	KE	ケニア	NO	ノルウェー	ZW	ジンバブエ
CI	コート・ジボアール	KG	キルギスタン	NZ	ニュージーランド		
CM	カメルーン	KP	朝鮮民主主義人民共和国	PL	ポーランド		
CN	中国	KR	大韓民国	PT	ポルトガル		
CU	キューバ	KZ	カザフスタン	RO	ルーマニア		
CZ	チェコ共和国	LI	リヒテンシュタイン	RU	ロシア連邦		
DE	ドイツ	LC	セントルシア	SD	スーダン		
DK	デンマーク	SK	スロバキア	SE	スウェーデン		
EE	エストニア	SL	シエラレオネ				

明 細 書

正準化データ作成装置、正準化データ作成方法及び正準化データ作成用記憶媒体

技術分野

本発明は、化学及び生化学分野の情報処理を行う処理装置及び処理方法に関し、特に、化合物を構成する各原子についての各種データから化合物の化学構造を一意的に特定する正準化データを作成する正準化データ作成装置及び正準化データ作成方法に関する。

また、本発明は、化学及び生化学分野の情報処理を行う処理プログラムが格納されたフレキシブルディスク、磁気テープ等の記憶媒体（コンピュータプログラム製品）に関し、特に、化合物を構成する各原子についての各種データから化合物の化学構造を一意的に特定する正準化データを作成するための処理プログラムが格納された記憶媒体に関する。

背景技術

従来より、化合物情報を収載した化合物データベースシステムやプログラム、及び化合物の反応情報を収載した反応データベースシステムやプログラムが開発されている。化合物データベースシステムやプログラムには、既存の化合物の物性や作用などの化合物情報が収載されており、化合物の構造をキーとして化合物情報にアクセスするものである。この化合物データベースを用いれば、化合物の特性や作用を効率良く参照することができる。また、反応データベースシステムやプログラムには、既存の化合物の反応情報が収載されており、化合物の構造をキーとして反応情報にアクセスするものである。この反応データベースを用いれば、合成化学の研究者が化合物の新規合成を行う際に、既存の化合物の反応情報の中から類似する反応情報を効率良く参照することができ、合成化学の研究では

必須である。

このような化合物データベースシステムとしては、例えば、米国MDL社の化合物管理システム“MACCS”がある。また、反応データベースシステムとしては、例えば、米国MDL社の総合化学情報管理システム“ISIS”や、反応情報管理システム“REACCS”がある。

ところで、従来の化合物／反応データベースシステムは、化合物の特性や作用の情報と共に、化合物の構造図をディスプレイに表示させる機能を有している。このように構造図を表示させることによって視覚的に優れたシステムの構築が可能となる。ところがこの化合物の構造図は、多くの記憶領域を占有する画像データ（ビットマップデータ）としてシステムに保存されるため、大容量の記憶装置が必要となり問題であった。

本発明は、このような問題を解決し、化合物／反応データベースシステムで利用した場合に、このシステムの記憶領域の使用量を大幅に削減させることのできる正準化データ作成装置、正準化データ作成方法及び正準化データ作成用記憶媒体（コンピュータプログラム製品）を提供することを目的とする。

発明の開示

本発明の正準化データ作成装置は、

化合物を構成する各原子についての固有データ及び原子間の結合対データの入力を受け付ける入力手段と、

前記入力手段で受け付けられた固有データ及び結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する正準化データ作成手段とを備えた正準化データ作成装置であって；

前記正準化データ作成手段は、

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える第1の処理部

と、

前記第1の処理部で各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える第2の処理部と、

前記第2の処理部で各原子に与えられた正準化番号に基づいて、前記正準化データを作成する第3の処理部とを備える。

上記本発明の正準化データ作成装置によれば、入力手段で受け付けられた各原子についての固有データ及び原子間の結合対データは正準化データ作成手段に与えられる。そして、正準化データ作成手段では、これらのデータに基づいて正準化データが作成される。

即ち、正準化データ作成手段では、まず、第1の処理部の処理を実行し、各原子についての固有データ及び原子間の結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類する。そして、クラスごとに異なるクラス番号を各原子に与える。次に、第2の処理部の処理を実行し、各原子に与えられたクラス番号及び原子間の結合対データに基づいて、化合物の構造と一意的に対応した正準化番号を各原子に与える。さらに、第3の処理部の処理を実行し、各原子に与えられた正準化番号及び各原子についての固有データに基づいて正準化データを作成する。

ここで、第1の処理部は、

各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記3種類の属性 (a_i , b_{ij} , d_{ij}) の中で、 a_i は入力番号 i の原子の種類記号であり、 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が j である結合の数であり、 d_{ij} は入力番号 i の原子から最短経路により、 j 個の結合を経て巡れる道筋の数であり；

第2の処理部は、

正準化番号を 1 から昇順に各原子に与える過程において、クラス番号の優先順位が最高である原子に正準化番号 1 を与え、以降正準化番号 n までが付与されているとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、クラス番号の優先順位が最高である原子に正準化番号 $n + 1$ を与えており；

第 3 の処理部は、

各原子に 3 種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一列に並べることによって正準化データを作成しており、

前記 3 種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号であることが好ましい。

本発明の正準化データ作成方法は、

化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する正準化データ作成方法であって；

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える第 1 のステップと、

前記第 1 のステップで各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える第 2 のステップと、

前記第 2 のステップで各原子に与えられた正準化番号に基づいて、前記正準化データを作成する第 3 のステップとを備える。

上記本発明の正準化データ作成方法によれば、上記第 1 のステップから第 3 の

ステップまでの処理によって、化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて正準化データが作成される。

即ち、まず、第1のステップの処理が実行され、各原子についての固有データ及び原子間の結合対データに基づいて、各原子が等価原子ごとに別のクラスに分類される。そして、クラスごとに異なるクラス番号が各原子に与えられる。次に、第2のステップの処理が実行され、各原子に与えられたクラス番号及び原子間の結合対データに基づいて、化合物の構造と一意的に対応した正準化番号が各原子に与えられる。さらに、第3のステップの処理が実行され、各原子に与えられた正準化番号及び各原子についての固有データに基づいて正準化データが作成される。

ここで、第1のステップは、

各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記3種類の属性 (a_i , b_{ij} , d_{ij}) の中で、 a_i は入力番号 i の原子の種類記号であり、 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が j である結合の数であり、 d_{ij} は入力番号 i の原子から最短経路により、 j 個の結合を経て巡れる道筋の数であり；

第2のステップは、

正準化番号を1から昇順に各原子に与える過程において、クラス番号の優先順位が最高である原子に正準化番号1を与え、以降正準化番号 n までが付与されているとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、クラス番号の優先順位が最高である原子に正準化番号 $n+1$ を与えており；

第3のステップは、

各原子に3種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一列に並べることによって正準化データを作成しており、

前記3種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号であることが好ましい。

本発明の正準化データ作成用コンピュータプログラム製品は、

化合物を構成する各原子についての固有データ及び原子間の結合対データの入力を受け付ける入力手段と、コンピュータ使用可能な媒体から情報を読み出す読出手段とを備える情報処理装置と共に用いるための正準化データ作成用コンピュータプログラム製品であって；

プログラムを記録するためのプログラム領域を有しており、かつ、前記固有データ及び結合対データに基づいて正準化データを効率的に作成させるための該媒体中に具現化されたコンピュータ読取可能なプログラムを有しているコンピュータ使用可能な媒体を前記コンピュータプログラム製品は備えており；

該コンピュータプログラム製品は、前記プログラム領域において、前記固有データ及び結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する、コンピュータ読取可能な正準化データ作成プログラムを有しており；

前記正準化データ作成プログラムは、

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える、コンピュータ読取可能な第1の処理ルーチンと、

前記第1の処理ルーチンで各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える、コンピュータ読取

可能な第2の処理ルーチンと、

前記第2の処理ルーチンで各原子に与えられた正準化番号に基づいて、前記正準化データを作成する、コンピュータ読取可能な第3の処理ルーチンとを備える。

上記本発明の正準化データ作成用コンピュータプログラム製品（正準化データ作成用記憶媒体）を所定の情報処理装置に収容して、プログラム領域に格納された正準化データ作成プログラムを読み取ることにより、正準化データ作成プログラムを情報処理装置で実行させることができる。正準化データ作成プログラムの起動によって、まず第1の処理ルーチンが実行されて、各原子についての固有データ及び原子間の結合対データに基づいて、各原子が等価原子ごとに別のクラスに分類される。そして、クラスごとに異なるクラス番号が各原子に与えられる。次に、第2の処理ルーチンが実行されて、各原子に与えられたクラス番号及び原子間の結合対データに基づいて、化合物の構造と一意的に対応した正準化番号が各原子に与えられる。さらに、第3の処理ルーチンが実行されて、各原子に与えられた正準化番号及び各原子についての固有データに基づいて正準化データが作成される。

ここで、前記第1の処理ルーチンは、

各原子に3種類の属性（ a_i ， b_{ij} ， d_{ij} ）を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記3種類の属性（ a_i ， b_{ij} ， d_{ij} ）の中で、 a_i は入力番号*i*の原子の種類記号であり、 b_{ij} は入力番号*i*の原子に隣接する結合のうち、その種類記号が*j*である結合の数であり、 d_{ij} は入力番号*i*の原子から最短経路により、*j*個の結合を経て巡れる道筋の数であり；

前記第2の処理ルーチンは、

正準化番号を1から昇順に各原子に与える過程において、クラス番号の優先順位が最高である原子に正準化番号1を与え、以降正準化番号*n*までが付与されて

いるとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、クラス番号の優先順位が最高である原子に正準化番号 $n + 1$ を与えており；

前記第 3 の処理ルーチンは、

各原子に 3 種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一行に並べることによって正準化データを作成しており、

前記 3 種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号であることが好ましい。

図面の簡単な説明

図 1 は、本発明の正準化データ作成装置の一例を示すブロック図である。

図 2 A は、結合表の原子テーブルのデータ内容の一例を示す図であり、図 2 B は、結合表の原子対テーブルのデータ内容の一例を示す図である。

図 3 は、化合物情報ファイルのデータ内容の一例を示す図である。

図 4 は、本発明の正準化データ作成装置の動作の概要を示す概略図である。

図 5 は、メインルーチンの処理の流れを示すフローチャートである。

図 6 は、構成原子分類ルーチンの処理の流れを示すフローチャートである。

図 7 A は、結合表の原子テーブルのデータ内容の一例を示す図であり、図 7 B は、結合表の原子対テーブルのデータ内容の一例を示す図である。

図 8 は、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子と入力番号との関係を示す図である。

図 9 A 及び図 9 B はそれぞれ、参照テーブルのデータ内容の一例を示す図であ

る。

図10は、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子に与えられた3種類の属性 (a_i , b_{ij} , d_{ij}) を示す図である。

図11A、図11B、図12、図13A、図13B、図14A及び図14Bはそれぞれ、参照テーブルのデータ内容の一例を示す図である。

図15A、図15B及び図15Cはそれぞれ、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子とクラス番号との関係を示す図である。

図16は、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子に与えられた属性 V_{ij}^1 を示す図である。

図17は、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子に与えられた属性 V_{ij}^2 を示す図である。

図18は、正準化番号付与ルーチンの処理の流れを示すフローチャートである。

図19は、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子と正準化番号との関係を示す図である。

図20は、正準化データ作成ルーチンの処理の流れを示すフローチャートである。

図21Aは、結合表の原子テーブルのデータ内容の一例を示す図であり、図21Bは、結合表の原子対テーブルのデータ内容の一例を示す図である。

図22は、正準化木構造データのデータ内容の一例を示す図である。

図23は、本発明の正準化データ作成用記憶媒体の一例の構成を示すブロック図である。

図24は、本発明にかかる正準化データ作成装置の一例を示すブロック図である。

図25は、本発明にかかる正準化データ作成装置の一例を示す斜視図である。

図26Aは、 C_{60} 分子の分子構造を示す図であり、図26Bは、 C_{60} 分子の正

準化データを示す図である。

発明を実施するための最良の形態

以下、本発明の好適な実施形態について添付図面を参照して説明する。

図1は、本発明の好適な実施形態に係る正準化データ作成装置1の構成を示すブロック図である。図1に示すように、正準化データ作成装置1は、化合物の分子構造図等の画像データ10aを記憶する画像メモリ10と、記号データ11a等を一時的に記憶する作業用メモリ11と、オペレーティングシステム(OS)21及び正準化データ作成プログラム22が記憶された主記憶装置20と、結合表ファイル31及び化合物情報ファイル33が記憶されたハードディスク装置30とを備えている。

また、正準化データ作成装置1は、化合物の分子構造図等を表示するディスプレイ40と、手書き図形の入力を受け付けるポインティングデバイスであるマウス50と、化学式等の記号データの入力を受け付けるキーボード60と、化合物の分子構造図等を出力するプリンタ70と、正準化データ作成プログラム22の実行等を制御するCPU80とを備えている。なお、ポインティングデバイスには、マウス50以外に、タブレット、ディジタイザ、ライトペンなどがあり、これらのいずれの装置をマウス50の代わりに備えてもよい。

正準化データ作成プログラム22は、化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて正準化データを作成するプログラム（正準化データ作成手段）である。この正準化データ作成プログラム22は、処理を統括するメインルーチン100と、化合物を構成する各原子にクラス番号を与える構成原子分類ルーチン（第1の処理ルーチン、第1の処理部）101とを備えている。また、正準化データ作成プログラム22は、クラス番号に基づいて各原子に正準化番号を与える正準化番号付与ルーチン（第2の処理ルーチン、第2の処理部）102と、各原子の正準化番号に基づいて正準化データを作成する

正準化データ作成ルーチン（第3の処理ルーチン、第3の処理部）103とを備えている。

ハードディスク装置30には、複数の結合表32を格納できる結合表ファイル31が設けられている。結合表32には化合物を構成する各原子についての固有データ及び原子間の結合対データが記録されており、正準化データ作成プログラム22は結合表32を介して、これらのデータにアクセスできる。

図2A及び図2Bに示すように、結合表32は、各原子についての固有データが記録された原子テーブル32aと、原子間の結合対データが記録された原子対テーブル32bとを備えている。具体的には、原子テーブル32aには、入力番号（原子の番号ともいう）、原子の二次元座標（X座標・Y座標）、元素名（一般に元素記号が用いられるが、原子番号などの数字であってもよい）、属性、原子数、及び結合数を書き込む欄が設けられており（図2A参照）、原子対テーブル32bには、結合原子対データ、結合種（例えば、単結合は1、二重結合は2とする）、及び構造（各分子が分子構造図の環状部に属するか、鎖状部に属するかを区別する欄）を書き込む欄が設けられている（図2B参照）。

ここで、入力番号は、化合物を構成する各原子をコンピュータで識別するための番号であり、図2Aの例では数字であるが、記号であっても良い。また、結合原子対データは、入力番号の組合せとして表現されるのが良い。

なお、正準化データを作成するためには、上記の原子テーブル32a及び原子対テーブル32bのすべてのデータが必要であるのではなく、固有データとして原子の番号及び元素名、結合対データとして結合原子対データ及び結合種があれば十分である。

また、ハードディスク装置30には、化合物番号とこの化合物に対応する正準化データ（カノニカルデータともいう）との関係を示す一覧表が記録された化合物情報ファイル33が格納されている。図3に示すように、化合物情報ファイル33は、化合物番号C₁～C₇の各化合物に対応した正準化データと、化合物C

1 ~ C₇ の各化合物についての参照データ（名称、文献、物性等）とを、化合物番号 C₁ ~ C₇ に対応した一覧表として記録したファイルである。このため、化合物番号 C₁ ~ C₇ をキーとして、化合物情報ファイル 33 にアクセスすれば、化合物番号 C₁ ~ C₇ の各化合物についての、正準化データ及び参照データを読み出すことができる。ここで、正準化データは、各化合物の化学構造を一意的に特定する複数の記号からなるデータである。

なお、構成原子分類ルーチン 101 が第 1 のステップに、正準化番号付与ルーチン 102 が第 2 のステップに、正準化データ作成ルーチン 103 が第 3 のステップにそれぞれ対応する。

次に、正準化データ作成装置 1 の動作の概要について説明する。図 4 に示すように、操作者はマウス 50 又はキーボード 60 を操作して、分子構造図の作成対象となる化合物の結合表 32 を結合表ファイル 31 内に作成することができる。

マウス 50 による入力、マウス 50 を用いてディスプレイ 40 上に化合物の分子構造図を手書き入力するもので、入力順に定まる各原子の入力番号が、主記憶装置 20 内に作成された結合表 32 の入力番号の欄に書き込まれる。さらに、この分子構造図 E₁ の各原子の結合関係を示す結合原子対データが、結合表 32 の結合原子対の欄に書き込まれる。このように、マウス 50 による入力では、化合物を特定する結合表 32 が手書きされた分子構造図 E₁ から作成される。

ここで、マウス 50 による手書き分子構造図 E₁ の入力について、より具体的に説明する。まず、操作者がマウス 50 をクリックすると、結合原子対を構成する一方の原子についてのデータが入力される。次に、操作者がマウス 50 を移動させて、マウス 50 を再びクリックすると、一方の原子と結合原子対を構成する他方の原子についてのデータが入力される。そして、この他方の原子についてのデータを入力するためのクリックは、続いて次のクリックがなされた時には、次の結合原子対の一方の原子を指定するクリックと見なされる。このように、連続した 2 回のクリックによって、一つの結合原子対を指定することができる。即ち、

操作者が原子を次々とずらしながら、結合原子対の指定を続けることにより、化合物を構成する全ての結合を入力することができる。

入力された原子についてのデータに対応して（例えば原子の入力順に 1, 2, 3, ... というように）、原子テーブル 3 2 a の入力番号の欄に原子の番号が書き込まれる。また、入力された原子の結合関係についてのデータは、原子対テーブル 3 2 b の結合原子対の欄に書き込まれる。さらに、操作者が原子の元素名を入力した場合には、それは原子テーブル 3 2 a の元素名の欄に書き込まれる。同様に、操作者が結合原子対を結ぶ結合の多重度を入力した場合には、それは原子対テーブル 3 2 b の結合種の欄に書き込まれる。元素名及び多重度は、それが書き込まれなかった時には、それぞれ炭素及び単結合と認識されるのがよい。なお、分子構造図及びそれに基づく結合表は、通常、水素原子を省略して作成される。

また、キーボード 6 0 による入力は、キーボード 6 0 を用いて所定の化合物に対応する結合表名を特定する記号列を入力するもので、入力された記号データ 1 1 a に基づいて、この結合表名によって特定される結合表 3 2 が結合表ファイル 3 1 から読み出される。

このように、マウス 5 0 及びキーボード 6 0 が入力手段 A を構成し、マウス 5 0 とキーボード 6 0 とのいずれを用いても結合表 3 2 が得られる。そして、正準化データ作成手段 B である正準化データ作成プログラム 2 2 が実行されて、結合表 3 2 の各データに基づいて正準化データ 3 4 が作成される。このように作成された正準化データ 3 4 は、化合物の参照データと共に化合物情報ファイル 3 3 に書き込まれて保存される。ここで、結合表 3 2 から正準化データ 3 4 を作成して保存するのは、結合表 3 2 のままで保存するよりも記憶領域を小さくできるからである。即ち、図 2 に示した結合表 3 2 に基づいて作成した正準化データ 3 4 は “1 % 1 % 1 - 2 % 3 % 5 % N / 6 % 7 /” であり、化合物の構造を非常に短い文字・数字・記号列で表すことができる。このように短い記号列を保存の対象と

すれば、記憶資源を有効に利用でき装置の小型軽量化に寄与することができる。

また、結合表 3 2 の結合原子対データに基づいて、二次元座標演算処理が行われることによって、各原子の二次元座標データを得ることができる。このように得られた二次元座標データから美的に優れた分子構造図 E₂ が作成される。作成された分子構造図 E₂ は、ディスプレイ 4 0 に表示することやプリンタ 7 0 から出力することができる。

なお、キーボード 6 0 による入力は、前記固有データ及び結合対データを主記憶装置 2 0 内に作成された結合表に直接書き込むことでも良い。また、イメージスキャナーやオプティカルカードリーダー（OCR）などの光学的に図形や文字を読み取る装置を本発明の入力装置として用いて、結合表データの人力の受け付けを行うことでも良い。

次に、本発明の好適な実施形態に係る正準化データ作成方法について説明する。この作成方法には、前述の正準化データ作成装置 1 が用いられる。まず、OS 2 1 の制御の下で、正準化データ作成プログラム 2 2 のメインルーチン 1 0 0 が起動される。

図 5 のフローチャートに示すように、メインルーチン 1 0 0 は、まず、構成原子分類ルーチン（第 1 の処理ルーチン）1 0 1 を呼び出して、化合物を構成する各原子にクラス番号を付与する（第 1 のステップ：S 1 0）。次に、正準化番号付与ルーチン（第 2 の処理ルーチン）1 0 2 を呼び出して、各原子に付与されたクラス番号に基づいて各原子に正準化番号を付与する（第 2 のステップ：S 2 0）。さらに、正準化データ作成ルーチン（第 3 の処理ルーチン）1 0 3 を呼び出して、各原子に付与された正準化番号に基づいて正準化データを作成する（第 3 のステップ：S 3 0）。このように作成された正準化データは、化合物情報ファイル 3 3 に書き込まれて保存される。

次に、S 1 0 で呼び出される構成原子分類ルーチン（第 1 の処理ルーチン）1 0 1 の処理（第 1 のステップ）について説明する。この処理は、化合物を構成す

る各原子を等価原子毎に別のクラスに分類して、属するクラスに対応するクラス番号を各原子に与える処理である。例えば、ベンゼンの6個の炭素原子は全て等価なので、全てに同一のクラス番号が与えられる。また、例えば、トルエンの7個の炭素原子は、5種類のクラス番号で表現される。即ち、トルエンのオルソ位の二つの炭素原子、さらにメタ位の二つの炭素原子はそれぞれ等価であり、同一のクラス番号が与えられる。

図6のフローチャートに示すように、まず、結合表32に基づいて化合物を構成する各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) をそれぞれ与える (S101)。ここで、属性 a_i は入力番号 i の原子の種類記号 (この例では原子番号) である。また、属性 b_{ij} は入力番号 i の原子に隣接する結合 (即ち、入力番号 i の原子を一方の原子とする結合) のうち、その種類記号 (この例では結合種は単結合を1、二重結合を2、三重結合を3、芳香結合を4とした) が j である結合の数 (ベクトル量) である。さらに、属性 d_{ij} は入力番号 i の原子から最短経路により j 個の結合を経て巡れる道筋の数 (ベクトル量) である。

次に、原子ごとに属性 (a_i , b_{ij} , d_{ij}) を並べて数字列 (この例では9桁の数字列) とし、この数字列が小さい順番にクラス番号 C_i^0 を与えて、各原子を複数のクラスに分類する (S102)。ここで与えられるクラス番号 C_i^0 は0次のクラス番号であり、S103以降のループ処理で1次のクラス番号 C_i^1 、2次のクラス番号 C_i^2 、…を順次求めていく。この0次の段階でクラスの数が増え、総原子数と等しければ、処理を終了してもよい。

次に、次数 n を1にする (S103)。そして、各原子に属性 V_{ij}^n を与える (S104)。属性 V_{ij}^n は入力番号 i の原子に結合し、次数 $n-1$ おいてクラス番号が j である原子の数である。さらに、原子ごとに属性 (a_i , b_{ij} , d_{ij} , V_{ij}^n) を並べて、この数字列が小さい順番にクラス番号 C_i^n を与えて、各原子を複数のクラスに分類する (S105)。そして、クラスの数 N_n が前のループでのクラスの数 $N_{(n-1)}$ と等しいか調べて、等しい場合には処理を終了する。

または、クラスの数 N_n が総原子数と等しいか調べて、等しい場合には処理を終了する(S106)。いずれも等しくない場合には、 n に1を加えて処理をS104に戻す(S107)。S102とS105の処理では、数字列が小さい順にクラス番号を与えているが、大きい順にクラス番号を与えてもよい。

次に、3, 5-ジメチルー2, 3, 4, 5-テトラヒドロピリジンを例にして構成原子分類ルーチン101のステップごとの処理を詳細に説明する。

まず、S101の処理が実行される。この処理を実行する際には、図7A及び図7Bに示すようなデータが既に結合表32に書き込まれており、この結合表32に書き込まれた各データに基づいて、各原子に3種類の属性(a_i , b_{ij} , d_{ij})を与える。ここで、この結合表32に記録された入力番号は、図8に示すように各原子が手書き入力された順番に与えられる任意の数字である。

属性 a_i は次のように求められる。上述したように属性 a_i は入力番号 i の原子の種類記号である。ここで、結合表32には各原子の元素記号が記録されており、これらの元素記号から原子の種類記号(原子番号)を求めることができる。従って、元素記号を結合表32から読み出すことによって、この元素記号に対応した属性 a_i を得ることができる。その結果、 a_1 , a_2 , $a_4 \sim a_8 = 6$ (炭素の原子番号)、 $a_3 = 7$ (窒素の原子番号)がそれぞれ得られる。

また、属性 b_{ij} は次のように求められる。上述したように属性 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が j である結合の数である。結合表32には各原子の結合種が記録されており、この結合種を結合表32から読み出すことによって属性 b_{ij} を得ることができる。その結果、 $b_{1j} = (3, 0, 0, 0)$ 、 $b_{2j} = (1, 1, 0, 0)$ 、 $b_{3j} = (1, 1, 0, 0)$ 、 $b_{4j} = (2, 0, 0, 0)$ 、 $b_{5j} = (3, 0, 0, 0)$ 、 $b_{6j} = (2, 0, 0, 0)$ 、 $b_{7j} = (1, 0, 0, 0)$ 、 $b_{8j} = (1, 0, 0, 0)$ がそれぞれ得られる。

具体的には、属性 b_{ij} は図9A及び図9Bに示す参照テーブル12を用いて求められる。参照テーブル12は2つの原子の結合関係を示す配列D(x, y)と

して構成されており、結合表 3 2 の結合原子対及び結合種のデータに基づいて作成される。ここで、 x は結合原子対データの 1 番目の原子の番号、 y は 2 番目の原子の番号であり、例えば $x = 2$ 、 $y = 3$ の座標位置には、結合種の 2 が丸で囲って示されている。即ち、結合原子対が示す配列要素に結合種 j が書き込まれて、参照テーブル 1 2 が作成されている。

この参照テーブル 1 2 を用いた属性 b_{ij} の抽出は次のように行われる。まず、参照テーブル 1 2 の配列要素のうち、 $x = 1$ 又は $y = 1$ を満たす配列要素（図 9 A において斜線で示した配列要素）を探索して、この配列要素に書き込まれたデータ（結合種） j を抽出する。その結果、 $D(1, 2) = 1$ 、 $D(1, 6) = 1$ 、 $D(1, 8) = 1$ が得られる。このように得られた 3 つの配列要素のデータ j は全て 1 なので、 $b_{11} = 3$ となる。また、データ j が 2 以上の配列要素は得られないので、 $b_{12} \sim b_{14} = 0$ となる。

次に、参照テーブル 1 2 の配列要素のうち、 $X = 2$ 又は $Y = 2$ を満たす配列要素（図 9 B において斜線で示した配列要素）を探索して、この配列要素に書き込まれたデータを抽出する。その結果、 $D(1, 2) = 1$ 、 $D(2, 3) = 2$ が得られる。このように得られた配列要素のデータ j は 1, 2 で、それぞれ 1 つなので、 $b_{21} = b_{22} = 1$ となる。また、データ j が 3 以上の配列要素は得られないので、 $b_{23} = b_{24} = 0$ となる。

$i = 3 \sim 8$ についても同様に処理することにより、図 1 0 に示す属性 b_{ij} ($i = 1 \sim 8$ 、 $j = 1 \sim 4$) が得られる。

さらに、属性 d_{ij} は次のように求められる。上述したように属性 d_{ij} は入力番号 i の原子から最短経路により j 個の結合を経て巡れる道筋の数である。即ち、図 8 の分子構造図に基づいて説明すると、入力番号 1 の原子から 1 個の結合を経て巡れる道筋は、（入力番号 1 ～ 入力番号 2）、（入力番号 1 ～ 入力番号 6）、（入力番号 1 ～ 入力番号 8）の計 3 本である。また、入力番号 1 の原子から 2 個の結合を経て巡れる道筋は、（入力番号 1 ～ 入力番号 2 ～ 入力番号 3）、（入力

番号1～入力番号6～入力番号5)の計2本である。

さらに、入力番号1の原子から最短経路により3個の結合を経て巡れる道筋は、(入力番号1～入力番号2～入力番号3～入力番号4)、(入力番号1～入力番号6～入力番号5～入力番号4)、(入力番号1～入力番号6～入力番号5～入力番号7)の計3本である。さらにまた、入力番号1の原子から最短経路により4個の結合を経て巡れる道筋は0本である。図8を見れば、入力番号1の原子から4個の結合を経て巡れる道筋は存在する。例えば、(入力番号1～入力番号2～入力番号3～入力番号4～入力番号5)の道筋である。しかしながら、入力番号1(出発点)から入力番号5(到着点)に至る道筋として、入力番号6を経由して2個の結合で巡れる道筋がある。従って、例示の道筋は出発点から到着点に至る最短経路ではない。以上の処理の結果、 $d_{1j} = (3, 2, 3, 0)$ が得られる。

同様の処理によって、 $d_{2j} = (2, 3, 2, 2)$ 、 $d_{3j} = (2, 2, 4, 0)$ 、 $d_{4j} = (2, 3, 2, 2)$ 、 $d_{5j} = (3, 2, 3, 0)$ 、 $d_{6j} = (2, 4, 2, 0)$ 、 $d_{7j} = (1, 2, 2, 3)$ 、 $d_{8j} = (1, 2, 2, 3)$ がそれぞれ得られる。

具体的には、属性 d_{ij} も属性 b_{ij} と同様に、参照テーブル12を参照して求められる。この参照テーブル12を参照した属性 d_{ij} の抽出は $i = 1$ 、 $i = 2$ 、…の順番で行われる。まず、属性 d_{1j} ($i = 1$) が抽出される。

属性 d_{1j} ($i = 1$) の抽出は、参照テーブル12の配列要素のうち、 $X = 1$ 又は $Y = 1$ を満たす配列要素(図11Aにおいて斜線で示した配列要素)を探索して、データが書き込まれた配列要素を抽出する。そして、抽出された配列要素に結合経路数として1を書き込む。その結果、 $D(1, 2)$ 、 $D(1, 6)$ 、 $D(1, 8)$ に結合経路数1が書き込まれる(図11Aでは結合経路数を三角形で囲んで示す)。

次に、結合経路数1が書き込まれた各配列要素の添字 $S = (1, 2)$ 、 $(1,$

6), (1, 8) を抽出する。これらの添字 S から前段の抽出処理で利用した 1 を除き、 $S = 2, 6, 8$ を得る。このようにして得られた $S = 2, 6, 8$ に基づいて、 $X = 2, 6, 8$ 又は $Y = 2, 6, 8$ を満たす配列要素 (図 1 1 B において斜線で示した配列要素) を探索し、データが書き込まれて且つ結合経路数が書き込まれていない配列要素を抽出する。そして、抽出された配列要素に結合経路数として 2 を書き込む。その結果、 $D(2, 3)$ 、 $D(5, 6)$ に結合経路数 2 が書き込まれる。

さらに、結合経路数 2 が書き込まれた各配列要素の添字 $S = (2, 3)$ 、 $(5, 6)$ を抽出する。これらの添字 S から前段の抽出処理で利用した 2, 6 を除き、 $S = 3, 5$ を得る。このようにして得られた $S = 3, 5$ に基づいて、 $X = 3, 5$ 又は $Y = 3, 5$ を満たす配列要素 (図 1 2 において斜線で示した配列要素) を探索し、データが書き込まれて且つ結合経路数が書き込まれていない配列要素を抽出する。そして、抽出された配列要素に結合経路数として 3 を書き込む。その結果、 $D(3, 4)$ 、 $D(4, 5)$ 、 $D(5, 7)$ に結合経路数 3 が書き込まれる。

以上の処理によって、全ての配列要素に結合経路数が書き込まれる。その結果、結合経路数 1 の配列要素が 3 個、結合経路数 2 の配列要素が 2 個、結合経路数 3 の配列要素が 3 個、結合経路数 4 の配列要素が 0 個となり、 $d_{ij} = (3, 2, 3, 0)$ が得られる。

次に、属性 $d_{2j} (i = 2)$ が抽出される。属性 $d_{2j} (i = 2)$ の抽出は、参照テーブル 1 2 の配列要素のうち、 $X = 2$ 又は $Y = 2$ を満たす配列要素 (図 1 3 A において斜線で示した配列要素) を探索して、データが書き込まれた配列要素を抽出する。そして、抽出された配列要素に結合経路数として 1 を書き込む。その結果、 $D(1, 2)$ 、 $D(2, 3)$ に結合経路数 1 が書き込まれる (図 1 3 A では結合経路数を三角形で囲んで示す)。

次に、結合経路数 1 が書き込まれた各配列要素の添字 $S = (1, 2)$ 、 $(2, 3)$ を抽出する。これらの添字 S から前段の抽出処理で利用した 2 を除き、 $S =$

1, 3を得る。このようにして得られた $S = 1, 3$ に基づいて、 $X = 1, 3$ 又は $Y = 1, 3$ を満たす配列要素（図 1 3 Bにおいて斜線で示した配列要素）を探索し、データが書き込まれて且つ結合経路数が書き込まれていない配列要素を抽出する。そして、抽出された配列要素に結合経路数として 2 を書き込む。その結果、 $D(1, 6)$ 、 $D(1, 8)$ 、 $D(3, 4)$ に結合経路数 2 が書き込まれる。

さらに、結合経路数 2 が書き込まれた各配列要素の添字 $S = (1, 6)$ 、 $(1, 8)$ 、 $(3, 4)$ を抽出する。これらの添字 S から前段の抽出処理で利用した 1, 3 を除き、 $S = 4, 6, 8$ を得る。このようにして得られた $S = 4, 6, 8$ に基づいて、 $X = 4, 6, 8$ 又は $Y = 4, 6, 8$ を満たす配列要素（図 1 4 Aにおいて斜線で示した配列要素）を探索し、データが書き込まれて且つ結合経路数が書き込まれていない配列要素を抽出する。そして、抽出された配列要素に結合経路数として 3 を書き込む。その結果、 $D(4, 5)$ 、 $D(5, 6)$ に結合経路数 3 が書き込まれる。

さらにまた、結合経路数 3 が書き込まれた各配列要素の添字 $S = (4, 5)$ 、 $(5, 6)$ を抽出する。これらの添字 S から前段の抽出処理で利用した 4, 6 を除き、 $S = 5, 5$ （即ち、 $S = 5$ が 2 重に適用される。）を得る。このようにして得られた $S = 5, 5$ に基づいて $X = 5$ 又は $Y = 5$ を満たす配列要素（図 1 4 Bにおいて斜線で示した配列要素）を探索して、データが書き込まれて且つ結合経路数が書き込まれていない配列要素を抽出する。そして、抽出された配列要素に結合経路数として 4 を書き込む。その結果、 $D(5, 7)$ に結合経路数 4 が 2 つ書き込まれる。

以上の処理によって、全ての配列要素に結合経路数が書き込まれる。その結果、結合経路数 1 の配列要素が 2 個、結合経路数 2 の配列要素が 3 個、結合経路数 3 の配列要素が 2 個、結合経路数 4 の配列要素が 2 個となり、 $d_{2j} = (2, 3, 2, 2)$ が得られる。

そして、属性 d_{ij} のその他 ($i = 3 \sim 8$) のそれぞれについて、同様の処理を

全ての配列要素に結合経路数が書き込まれるまで続けると、図10に示す d_{ij} ($i = 1 \sim 8$, $j = 1 \sim 4$) が得られる。以上説明したS101の処理によって、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを構成する各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) が与えられる。

次に、S102の処理が実行される。上述したように、S102では原子ごとに属性 (a_i , b_{ij} , d_{ij}) を並べて9桁の数字列とし、この数字列が小さい順番にクラス番号 C_i^0 を与えて、各原子を複数のクラスに分類している。ここで与えられるクラス番号 C_i^0 は入力番号 i の原子の0次のクラス番号である。

S102の処理を具体的に説明すると、入力番号1の原子の数字列は“630003230”であり、入力番号2の原子の数字列は“611002322”である。以下順番に、“711002240”、“620002322”、“630003230”、“620002420”、“610001223”、“610001223”となる。

その結果、入力番号7及び8の原子の数字列が最小となり、これらの原子にクラス番号 $C_7^0 = C_8^0 = 1$ が与えられる。以下、数字列が小さい順に、入力番号2の原子にクラス番号 $C_2^0 = 2$ が与えられて、入力番号4の原子にクラス番号 $C_4^0 = 3$ が与えられる。また、入力番号6の原子にクラス番号 $C_6^0 = 4$ が与えられて、入力番号1及び5の原子にクラス番号 $C_1^0 = C_5^0 = 5$ が与えられる。さらに、入力番号3の原子にクラス番号 $C_3^0 = 6$ が与えられる(図15A参照)。このようにして各原子は6つのクラスに分類されて、クラスの数 N_0 は6となる。

次に、S103の処理が実行されて、次元 n を1にする。

さらに、S104の処理が実行される。上述したように、S104では各原子に属性 $V_{ij}^{(n=1)}$ を与えている。ここで、属性 $V_{ij}^{(n=1)}$ は入力番号 i の原子に結合し、クラス番号が j である原子の数である。即ち、図15Aの分子構造図に基づいて説明すると、入力番号1の原子に結合した原子の入力番号は2, 6, 8であ

り、これらの原子のクラス番号は $C_2^0 = 2$ 、 $C_6^0 = 4$ 、 $C_8^0 = 1$ である。その結果、 $j = 1, 2, 4$ の属性 V_{1j}^1 に 1 が書き込まれて、 $V_{1j}^1 = (1, 1, 0, 1, 0, 0)$ が得られる。

また、入力番号 2 の原子に結合した原子の入力番号は 1, 3 であり、これらの原子のクラス番号は $C_1^0 = 5$ 、 $C_3^0 = 6$ である。その結果、 $j = 5, 6$ の属性 V_{2j}^1 に 1 が書き込まれて、 $V_{2j}^1 = (0, 0, 0, 0, 1, 1)$ が得られる。入力番号 3～8 の原子についても同様に処理することにより、 $V_{3j}^1 = (0, 1, 1, 0, 0, 0)$ 、 $V_{4j}^1 = (0, 0, 0, 0, 1, 1)$ 、 $V_{5j}^1 = (1, 0, 1, 1, 0, 0)$ 、 $V_{6j}^1 = (0, 0, 0, 0, 2, 0)$ 、 $V_{7j}^1 = (0, 0, 0, 0, 1, 0)$ 、 $V_{8j}^1 = (0, 0, 0, 0, 1, 0)$ がそれぞれ得られる。

具体的には、属性 V_{ij}^1 は図 9 A 及び図 9 B に示す参照テーブル 12 を用いて求められる。この参照テーブル 12 を用いた属性 V_{ij}^1 の抽出は $i = 1$ 、 $i = 2$ 、…の順番で行われる。まず、属性 V_{1j}^1 ($i = 1$) が抽出される。属性 V_{1j}^1

($i = 1$) の抽出は、参照テーブル 12 の配列要素のうち、 $x = 1$ 又は $y = 1$ を満たす配列要素 (図 9 A において斜線で示した配列要素) を探索して、データが書き込まれた配列要素の添字 $S = (1, 2)$ 、 $(1, 6)$ 、 $(1, 8)$ を抽出する。これらの添字 S から $i = 1$ を除き、 $S = 2, 6, 8$ を得る。このようにして得られた S の値を 0 次のクラス番号 C_i^0 に代入して、 $C_2^0 = 2$ 、 $C_6^0 = 4$ 、 $C_8^0 = 1$ を得る。そして、 $j = 1, 2, 4$ の属性 V_{1j}^1 に 1 を書き込むことにより、 $V_{1j}^1 = (1, 1, 0, 1, 0, 0)$ が求められる。

次に、属性 V_{2j}^1 ($i = 2$) が抽出される。属性 V_{2j}^1 ($i = 2$) の抽出は、参照テーブル 12 の配列要素のうち、 $X = 2$ 又は $Y = 2$ を満たす配列要素 (図 9 B において斜線で示した配列要素) を探索して、データが書き込まれた配列要素の添字 $S = (1, 2)$ 、 $(2, 3)$ を抽出する。これらの添字 S から $i = 2$ を除き、 $S = 1, 3$ を得る。このようにして得られた S の値を 0 次のクラス番号 C_i^0 に代入して、 $C_1^0 = 5$ 、 $C_3^0 = 6$ を得る。そして、 $j = 5, 6$ の属性 V_{2j}^1

1 に 1 を書き込むことにより、 $V_{2,j}^1 = (0, 0, 0, 0, 1, 1)$ が求められる。

$i = 3 \sim 8$ についても同様に処理することにより、図 16 に示す属性 $V_{i,j}^1$ ($i = 1 \sim 8$ 、 $j = 1 \sim 6$) が得られる。

次に、S 105 の処理が実行される。上述したように、S 105 では原子ごとに属性 $(C_i^{n-1}, V_{i,j}^n)$ を並べて、この数字列が小さい順番にクラス番号 C_i^n を与えて、各原子を複数のクラスに分類している。

具体的には、入力番号 1 の原子の数字列は “5 1 1 0 1 0 0” であり、入力番号 2 の原子の数字列は “2 0 0 0 0 1 1” である。以下順番に、“6 0 1 1 0 0 0”、“3 0 0 0 0 1 1”、“5 1 0 1 1 0 0”、“4 0 0 0 0 2 0”、“1 0 0 0 0 1 0”、“1 0 0 0 0 1 0” となる。

その結果、入力番号 7 及び 8 の原子の数字列が最小となり、これらの原子にクラス番号 $C_7^1 = C_8^1 = 1$ が与えられる。以下、数字列が小さい順に、入力番号 2 の原子にクラス番号 $C_2^1 = 2$ が与えられて、入力番号 4 の原子にクラス番号 $C_4^1 = 3$ が与えられる。さらに、入力番号 6 の原子にクラス番号 $C_6^1 = 4$ が与えられて、入力番号 5 の原子にクラス番号 $C_5^1 = 5$ が与えられる。さらにまた、入力番号 1 の原子にクラス番号 $C_1^1 = 6$ が与えられて、入力番号 3 の原子にクラス番号 $C_3^1 = 7$ が与えられる。このようにして各原子は 7 つのクラスに分類されて、クラスの数 N_1 は 7 となる。

次に、S 106 の処理を実行して、クラスの数 N_n が $N_{(n-1)}$ と等しいか調べて、等しい場合には処理を終了する。また、クラスの数 N_n が総原子数と等しいか調べて、等しい場合には処理を終了する。ここでは、クラスの数 N_1 が 7 で、クラスの数 N_0 が 6 なので、 N_1 と N_0 とは等しくない。また、総原子数は 8 なので、クラスの数 N_1 と総原子数とは等しくない。このようにいずれも等しくないので、S 107 の処理を実行して n を 2 とする。

さらに、S 104 に戻って各原子に属性 $V_{i,j}^2$ を与える。その結果、図 17 に

示すように、 $V_{1j}^2 = (1, 1, 0, 1, 0, 0, 0)$ 、 $V_{2j}^2 = (0, 0, 0, 0, 0, 1, 1)$ 、 $V_{3j}^2 = (0, 1, 1, 0, 0, 0, 0)$ 、 $V_{4j}^2 = (0, 0, 0, 0, 1, 0, 1)$ 、 $V_{5j}^2 = (1, 0, 1, 1, 0, 0, 0)$ 、 $V_{6j}^2 = (0, 0, 0, 0, 1, 1, 0)$ 、 $V_{7j}^2 = (0, 0, 0, 0, 1, 0, 0)$ 、 $V_{8j}^2 = (0, 0, 0, 0, 0, 1, 0)$ が得られる。

そして、S 1 0 5 の処理を実行して、各原子にクラス番号 C_i^2 を与える。その結果、図 1 5 C に示すように、 $C_1^2 = 7$ 、 $C_2^2 = 3$ 、 $C_3^2 = 8$ 、 $C_4^2 = 4$ 、 $C_5^2 = 6$ 、 $C_6^2 = 5$ 、 $C_7^2 = 2$ 、 $C_8^2 = 1$ が得られる。このようにして各原子は 8 つのクラスに分類されて、クラスの数 N_2 は 8 となる。クラスの数 $N_2 = 8$ は総原子数と等しいので、S 1 0 6 の判定によって処理を終了する。

次に、図 1 8 のフローチャートを用いて、図 5 の S 2 0 で呼び出される正準化番号付与ルーチン（第 2 の処理ルーチン）1 0 2 の処理（第 2 のステップ）について説明する。ここで正準化番号とは、化合物の構造によって一意的に定まる各原子の番号である。即ち、分子構造図を手書き入力することによって与えられる入力番号は、入力する順番が異なることによって変わる任意的な番号である。これに対して正準化データ 3 4 は、化合物の構造にのみ依存した一意的なデータでなければならない。このため、任意的な入力番号から一意的な正準化データ 3 4 を直接作成することは困難である。そこで、正準化データ作成プログラム 2 2 では、入力番号を一旦正準化番号に変換して、この一意的な正準化番号に基づいて正準化データ 3 4 を作成することにより、円滑な正準化データ 3 4 の作成を可能にしている。

正準化番号付与ルーチン 1 0 2 の処理は、まず、変数 k に 1 を与える（S 2 0 1）。次に、構成原子分類ルーチン 1 0 1 で得られた最終クラス番号 C_i^1 を調べて、最大の原子に正準化番号 k （ここでは $k = 1$ ）を与える（S 2 0 2）。クラス番号が最大の原子が複数個ある場合には、これらの原子の中から任意の原子を選び、この原子に正準化番号 k を与える。そして、全ての原子に正準化番号が

与えられた後に処理を終了する (S 2 0 3)。

次に、変数 k に 1 を加えて (S 2 0 4)、正準化番号の決まった原子 (以下、既決原子という) の中から、まだ正準化番号が決まっていない原子 (以下、未決原子という) が結合している既決原子を抽出する (S 2 0 5)。そして、抽出された既決原子が複数あるか判定して (S 2 0 6)、抽出された既決原子が複数ある場合には、これらの既決原子の中で正準化番号が最小の既決原子を選択する (S 2 0 7)。そして、選択された既決原子に結合している未決原子の中からクラス番号 C_i が最大の未決原子を抽出して、この未決原子の正準化番号を k とする (S 2 0 8)。なお、クラス番号 C_i が最大の既決原子が複数ある場合には、これらの既決原子の中から任意に選択する。

また、S 2 0 6 で既決原子が 1 つであると判定された場合には、この既決原子に結合している未決原子の中からクラス番号 C_i が最大の未決原子を選択して、この未決原子に正準化番号 k を与える (S 2 0 9)。S 2 0 8 及び S 2 0 9 の処理が終了した後に S 2 0 3 に処理を戻し、全ての原子に正準化番号が与えられるまで、S 2 0 3 ~ S 2 0 9 のループを繰り返す。

次に、正準化番号付与ルーチン 1 0 2 の処理について、3, 5-ジメチル-2, 3, 4, 5-テトラヒドロピリジンを用いた具体例を説明する。まず、S 2 0 1 の処理で変数 k に 1 を与えて、次に S 2 0 2 の処理を行う。S 2 0 2 の処理では、入力番号 3 の原子が $C_i = 8$ で最大なので、入力番号 3 の原子に正準化番号 $k = 1$ を与える。次に、S 2 0 4 の処理で変数 k を 2 にして、S 2 0 5 の処理で入力番号 3 の原子を既決原子として抽出する。

このように抽出された既決原子は 1 つだけなので、次に S 2 0 9 の処理を行う。入力番号 3 の原子に結合している未決原子は入力番号 2, 4 の原子なので、これらの原子の中からクラス番号 C_i が最大の原子を選択する。即ち、入力番号 2 の原子のクラス番号は $C_i = 3$ で、入力番号 4 の原子のクラス番号は $C_i = 4$ である。このため、入力番号 4 の原子を選択して、この原子に正準化番号 $k =$

2を与える。

次に、S 2 0 4 の処理に戻って変数 k を 3 にして、S 2 0 5 の処理で入力番号 3, 4 の原子を既決原子として抽出する。このように抽出された既決原子は複数あるので、次に S 2 0 7 の処理を行い、抽出された既決原子の中から正準化番号が最小の原子を選択する。即ち、入力番号 3 の原子の正準化番号は 1 で、入力番号 4 の原子の正準化番号は 2 である。このため、入力番号 3 の原子を選択する。そして、S 2 0 8 の処理を行い、入力番号 3 の原子に結合した入力番号 2 の原子に正準化番号 $k = 3$ を与える。

さらに、S 2 0 4 の処理に戻って変数 k を 4 にして、S 2 0 5 の処理で入力番号 2, 4 の原子を既決原子として抽出する。このように抽出された既決原子は複数あるので、次に S 2 0 7 の処理を行い、抽出された既決原子の中から正準化番号が最小の原子を選択する。即ち、入力番号 2 の原子の正準化番号は 3 で、入力番号 4 の原子の正準化番号は 2 である。このため、入力番号 4 の原子を選択する。そして、S 2 0 8 の処理を行い、入力番号 4 の原子に結合した入力番号 5 の原子に正準化番号 $k = 4$ を与える。

同様の処理を繰り返すことにより、入力番号 1 の原子に正準化番号 5 が、入力番号 6 の原子に正準化番号 6 がそれぞれ与えられる。また、入力番号 7 の原子に正準化番号 7 が、入力番号 8 の原子に正準化番号 8 がそれぞれ与えられる。

その後、S 2 0 3 の処理を行い、この段階では全ての原子の正準化番号が求められているので処理を終了する。その結果、図 1 9 に示すような正準化番号が得られる。

本発明の構成原子分類ルーチン 1 0 1 と正準化番号付与ルーチン 1 0 2 は、原子への番号付与 (S 1 0 2, S 1 0 5, S 2 0 2, S 2 0 8, S 2 0 9) という処理と、原子の選択 (S 2 0 2, S 2 0 7, S 2 0 8, S 2 0 9) という判断とを含んでいる。そして、番号付与にあたって昇順で行うか降順で行うか、また原子の選択にあたって数字の大小の大きいものを優先するか小さいものを優先する

かは、化合物の構造と一意的に対応する正準化データを作成するという課題が達成される範囲で、プログラム作成者の自由な選択である（負の数学を用いるという選択を含めて）。

従って、本発明における「クラス番号の優先順位が最高である」とは、上述の意味であって、必ずしも数字の大きな方を選択することを意味しない。但し、正準化番号を付与する過程で、本例のように既決原子から順に右左右左・・・と交互に番号が付与されるアルゴリズムを採用するのが好ましい。

さらに、本発明の第2の処理部、第2のステップ又は第2の処理ルーチンにおいて、正準化番号を1から各原子に昇順に番号を付けるやり方は、等差数列的（ $\{p + q(n - 1) \mid n = 1, 2, 3 \cdots n_{\max}\}$ ）であればよく、「1」は基準となる数学（等差数列の初項 p ）という意味であり、1そのものである必要はない。また、公差 q も1でなくてもよい。

さらにまた、本発明の第2の処理部、第2のステップ又は第2の処理ルーチンにおいて、正準化番号を初項 p （通常、総原子数とされる）から降順（負の公差 q ）に各原子に与えることもできる。この場合には、正準化番号 n の次には、既に正準化番号が与えられている原子で、且つ正準化番号がまだ与えられていない原子が結合している原子の中で、正準化番号が最大である原子を選び、その原子に結合している原子で、且つ正準化番号がまだ与えられていない原子の中で、クラス番号の優先順位が最高である原子に次の正準化番号 $n + q$ （ q は負なので、実際には $n - 1$ のような値である）が与えられる。降順の場合には、さらに、属性 P_i の定義は、正準化番号 i の原子に結合し且つ正準化番号が最大の原子の正準化番号、と改められるのがよい。

次に、図20のフローチャートを用いて、S30で呼び出される正準化データ作成ルーチン（第3の処理ルーチン）103の処理（第3のステップ）について説明する。この処理は、まず、図21に示すように入力番号を正準化番号に置き換えて、結合表32を書き替える（S301）。そして、この結合表32に基づ

いて、各原子に対して3種類のデータ (P_i , T_i , S_i) を求める (S 3 0 2)。ここで、 P_i は正準化番号 i ($i > 1$) の原子に結合して、番号が最小である原子の正準化番号である。また、 T_i は正準化番号 i ($i > 1$) の原子と正準化番号 P_i の原子との結合の種類記号 (この例では単結合は一、二重結合は=、三重結合は#、芳香結合は%など) である。さらに、 S_i は正準化番号 i ($i > 0$) の原子の種類記号 (この例では元素記号) である。

具体的には、まず、正準化番号 1 の原子の元素記号について、原子テーブル 3 2 a を参照して調べる。その結果、 $S_1 = \text{"N"}$ が得られる。次に、正準化番号 2 の原子と結合した原子について、原子対テーブル 3 2 b を参照して調べる。その結果、正準化番号 1, 4 の原子が得られる。これらの原子の中で最小の正準化番号は 1 なので、 $P_2 = 1$ となる。そして、正準化番号 2 の原子と正準化番号 1 の原子との結合は単結合なので、 $T_2 = \text{"-"}$ となる。さらに、原子テーブル 3 2 a を参照することにより、 $S_2 = \text{"C"}$ が得られる。

次に、正準化番号 3 の原子と結合した原子について、原子対テーブル 3 2 b を参照して調べる。その結果、正準化番号 1, 5 の原子が得られる。これらの原子の中で最小の正準化番号は 1 なので、 $P_3 = 1$ となる。そして、正準化番号 3 の原子と正準化番号 1 の原子との結合は二重結合なので、 $T_3 = \text{"="}$ となる。さらに、原子テーブル 3 2 a を参照することにより、 $S_3 = \text{"C"}$ が得られる。以下同様に処理を行うことにより、 $P_4 = 2$ 、 $P_5 = 3$ 、 $P_6 = 4$ 、 $P_7 = 4$ 、 $P_8 = 5$ 、 $T_4 \sim T_8 = \text{"-"}$ 、 $S_4 \sim S_8 = \text{"C"}$ が得られる。

次に、S 3 0 2 の処理で T_i を求めた際に参照されなかった結合原子対を抽出する (S 3 0 3)。この処理は原子対テーブル 3 2 b を参照して行う。その結果、正準化番号 5 の原子と正準化番号 6 の原子との結合原子対が抽出される。そして、抽出された結合原子対に対して3種類のデータ (R^1_j , R^2_j , H_j) を求める (S 3 0 4)。ここで、 R^1_j , R^2_j はその結合を構成する2つの原子の正準化番号である。また、 H_j はその結合の種類記号 (この例では T_i と同じもの

を用いる)である。なお、 R^1_j と R^2_j とは、 $R^1_j > R^2_j$ の関係を満たすものとする。また、別の結合原子対 (R^1_k, R^2_k) とは、 $R^1_j \leq R^1_k$ の関係を満たすか、 $R^1_j = R^1_k$ で且つ $R^2_j < R^2_k$ の関係を満たすものとする。

以上の処理によって、図 2 2 に示す正準化木構造データが作成できる。

次に、S 3 0 2 及び S 3 0 4 の処理で求めた各データを一系列に並べて、正準化データを作成する (S 3 0 5)。即ち、原子の種類記号及び結合の種類記号と異なる区切り記号 F を定義して、S 3 0 2 及び S 3 0 4 の処理で求めた各データを以下のように並べる。

$S_1, P_2, T_2, S_2, P_3, T_3, S_3, P_4, T_4, S_4, \dots, P_N, T_N, S_N, F, R^1_1, H_1, R^2_1, F, R^1_2, H_2, R^2_2, \dots, F, R^1_M, H_M, R^2_M, F$

ここで、N は総原子数であり、M は S 3 0 4 の抽出された結合原子対の総数である。

このようにして得られたデータ列は、化合物の構造と一意的に対応する正準化データである。具体的には、区切り記号 F を “/” として、得られたデータを所定の順番に並べると、

“N 1 = C 1 = C 2 - C 3 - C 4 - C 4 - C 5 - C / 5 - 6 /”

が得られる。そして、この正準化データは化合物情報ファイル 3 3 に書き込まれて保存される (S 3 0 6)。その後、処理は終了する。

次に、本発明の好適な実施形態に係る正準化データ作成用コンピュータプログラム製品 (記憶媒体) について説明する。

図 2 3 は、本発明に実施形態に係る正準化データ作成用コンピュータプログラム製品 (記憶媒体) 2 の構成を示すブロック図である。図 2 3 に示すように、正準化データ作成用記憶媒体 2 は、ファイルを格納するファイル領域 2 b と、プログラムを格納するプログラム領域 2 a とを備えている。そして、ファイル領域 2

bには、前述の結合表ファイル31と、化合物情報ファイル33とが格納されている。

また、プログラム領域2aには、化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて正準化データを作成するための前述の正準化データ作成プログラム22が格納されている。この正準化データ作成プログラム22は、処理を統括するメインルーチン100と、化合物を構成する各原子にクラス番号を与える構成原子分類ルーチン（第1の処理ルーチン）101と、クラス番号に基づいて各原子に正準化番号を与える正準化番号付与ルーチン（第2の処理ルーチン）102と、各原子の正準化番号に基づいて正準化データを作成する正準化データ作成ルーチン（第3の処理ルーチン）103とを備えている。

正準化データ作成用記憶媒体2としては、例えば、フレキシブルディスクやCD-ROMなどの円盤型記憶媒体が用いられる。また、磁気テープなどのテープ型記憶媒体を用いてもよい。

本実施形態の正準化データ作成用記憶媒体2に格納された正準化データ作成プログラム22は、例えば図24及び図25に示す情報処理装置（正準化データ作成装置）1で実行させることができる。

図24は本実施形態の情報処理装置1の構成を示すブロック図であり、図25はその斜視図である。情報処理装置1は、正準化データ作成用記憶媒体2に格納された正準化データ作成プログラム22を読み取るための媒体ドライブ装置3を備えており、この媒体ドライブ装置3に正準化データ作成用記憶媒体2を収納することができる。そして、この収納によって、正準化データ作成用記憶媒体2に格納された情報（正準化データ作成プログラム22、結合表ファイル31及び化合物情報ファイル33）に対して情報処理装置1からアクセス可能となる。それによって、プログラム領域2aに格納された正準化データ作成プログラム22が情報処理装置1で実行できるようになる。

この情報処理装置1の構成は、次の通りである。すなわち、情報処理装置1は、

上述の媒体ドライブ装置 3 と、オペレーティングシステム (OS) 21 が記憶された主記憶装置 20 と、化合物の分子構造図等を示す画像データを記憶する画像メモリ 10 と、記号データ 11a などを一時的に記憶する作業用メモリ 11 と、化合物の分子構造図等を表示するディスプレイ 40 とを備えている。さらに、情報処理装置 1 は、手書き図形の入力を受け付けるポインティングデバイスであるマウス 50 と、化学式等の記号データの入力を受け付けるキーボード 60 と、化合物の分子構造図等を入力するプリンタ 70 と、正準化データ作成プログラム 22 の実行等を制御する CPU 80 とを備えている。

媒体ドライブ装置 3 としては、正準化データ作成用記憶媒体 2 に対応して、フレキシブルディスクドライブ装置、CD-ROM ドライブ装置、或いは磁気テープドライブ装置などが用いられる。

次に、正準化データ作成用記憶媒体 2 に格納された正準化データ作成プログラム 22 を情報処理装置 1 で実行させた場合の処理の流れについて説明する。図 24 に示すように、正準化データ作成用記憶媒体 2 が媒体ドライブ装置 3 に挿入されると、正準化データ作成用記憶媒体 2 のプログラム領域 2a に格納された正準化データ作成プログラム 22 が媒体ドライブ装置 3 によって読み出される。そして、媒体ドライブ装置 3 で読み出された正準化データ作成プログラム 22 は、主記憶装置 20 に転送され、主記憶装置 20 に記憶される。その後、キーボード 60 等を用いて操作者が動作開始の命令を入力すると、OS 21 の制御の下で正準化データ作成プログラム 22 のメインルーチン 100 が起動される。

そしてその後、前述の構成原子分類ルーチン (第 1 の処理ルーチン) 101、正準化番号付与ルーチン (第 2 の処理ルーチン) 102 及び正準化データ作成ルーチン (第 3 の処理ルーチン) 103 が前述の通りに実行され、化合物を一意的に特定する正準化データが短時間で得られる。

以上、本発明の正準化データ作成装置、作成方法及び作成用コンピュータプログラム製品の好適な実施形態について説明したが、本発明は上記の実施形態に限

定されるものではなく、本発明の趣旨から逸脱しない範囲内において、例えば以下のように変更することも可能である。

(1) 上記実施形態では、正準化データとして原子の種類記号 S_i を含めたデータ列を用いているが、最も出現頻度が高い原子の種類記号（通常は炭素の C）をデータ列から除いてもよい。即ち、上述した正準化データから炭素 C の記号を省略することにより、

“N 1 - 1 = 2 - 3 - 4 - 4 - 5 - / 5 - 6 /”

が得られる。このようにしてデータ列を短くすることにより、化合物情報ファイル 33 に書き込まれるデータ量を削減することができる。

(2) 正準化番号付与ルーチン 102 では、S 209 の処理でクラス番号 C_i が最大の未決原子が複数選択された場合に、次の処理を追加してもよい。

(a) クラス番号 C_i が最大の未決原子が環状構造部分に属していない場合には、複数の未決原子の中から任意の未決原子を選択して、この未決原子の正準化番号を k とする。その後処理を S 203 に戻す。

(b) クラス番号 C_i が最大の未決原子が環状構造部分に属している場合には、S 209 で選択された未決原子（以下、候補原子という）とこれらの候補原子に結合した既決原子との結合を切断した構造について、各候補原子に対して次のベクトル量を定義する。

m_{ik} : 候補原子 i と、正準化番号 k である原子間の最小結合数

予め、この属性について優先順位を定めておき、最も優先順位の高い原子 i を選択して、その原子の正準化番号を k とする。その後処理を S 203 に戻す。

ここで、原子の属性値に基づく原子の選択基準を例示する。まず、スカラー量についてはその大小による。また、ベクトル量については 2 つのベクトル i, k の要素が V_{ij}, V_{kj} のとき、 $V_{ij} \neq V_{kj}$ である要素の中で最小の j における大小を優先順位の判定基準とする。このような判定基準を用いることにより、属性 $b_{ij}, d_{ij}, V_{ij}, m_{ij}$ の優先順位を定めることができる。また複数の属性によ

って優先順位が定まる場合には、予め、属性間にも優先順位を定めておき、優先順位の高い属性での判定を優先するのが好ましい。

なお、上記の本発明にかかる正準化データ作成装置（正準化データ作成用コンピュータプログラム製品）を用いて本発明にかかる正準化データ作成方法によって図26Aに示すC₆₀分子の正準化データを求めたところ、C₆₀分子の構造を一意的に特定する正準化データ（図26B）を僅か1.5秒で得ることができた。これに対して、原子を等価原子毎に分類する処理過程を経ていないモルガン・アルゴリズム（H.L.Morgan, J.Chem.Doc., 5(2), 107(1965)）によって、同一性能の情報処理装置を用いてC₆₀分子の正準化データを求めたところ、正準化データを得るのに550秒を要した。従って、上記本発明にかかる正準化データ作成装置、作成方法及び作成用コンピュータプログラム製品を採用すれば、正準化データの作成に要する情報処理の速度及び正確性が大幅に向上する。

産業上の利用可能性

以上詳細に説明したように、本発明の正準化データ作成装置においては、入力手段で受け付けられた各原子についての固有データ及び原子間の結合対データは正準化データ作成手段に与えられる。そして、正準化データ作成手段では、これらのデータに基づいて正準化データが短時間でかつ正確に作成される。また、本発明の正準化データ作成方法によれば、化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて正準化データが短時間でかつ正確に作成される。さらに、本発明の正準化データ作成用コンピュータプログラム製品（記憶媒体）によれば、プログラムエリアに格納された正準化データ作成プログラムの実行によって、化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて正準化データが短時間でかつ正確に作成される。

また、このように本発明の正準化データ作成装置、正準化データ作成方法及び正準化データ作成用コンピュータプログラム製品で作成された正準化データは、

非常に短い文字・数字・記号列であり、少ない記憶領域で正準化データを保存することができる。このため、本発明の正準化データ作成装置、作成方法及び作成用コンピュータプログラム製品を化合物／反応データベースシステム等で利用すれば、化合物／反応データベースシステムの記憶領域の使用量を大幅に削減させることができる。

請 求 の 範 囲

1. 化合物を構成する各原子についての固有データ及び原子間の結合対データの入力を受け付ける入力手段と、

前記入力手段で受け付けられた固有データ及び結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する正準化データ作成手段とを備えた正準化データ作成装置であって；

前記正準化データ作成手段は、

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える第1の処理部と、

前記第1の処理部で各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える第2の処理部と、

前記第2の処理部で各原子に与えられた正準化番号に基づいて、前記正準化データを作成する第3の処理部とを備える；

正準化データ作成装置。

2. 前記第1の処理部は、

各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記3種類の属性 (a_i , b_{ij} , d_{ij}) の中で、 a_i は入力番号 i の原子の種類記号であり、 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が j である結合の数であり、 d_{ij} は入力番号 i の原子から最短経路により、 j 個の結合を経て巡れる道筋の数であり；

前記第2の処理部は、

正準化番号を1から昇順に各原子に与える過程において、前記クラス番号の優

先順位が最高である原子に正準化番号 1 を与え、以降正準化番号 n までが付与されているとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、前記クラス番号の優先順位が最高である原子に正準化番号 $n + 1$ を与えており；

前記第 3 の処理部は、

各原子に 3 種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一行に並べることによって前記正準化データを作成しており、

前記 3 種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号である；

請求項 1 に記載の正準化データ作成装置。

3. 化合物を構成する各原子についての固有データ及び原子間の結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する正準化データ作成方法であって；

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える第 1 のステップと、

前記第 1 のステップで各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える第 2 のステップと、

前記第 2 のステップで各原子に与えられた正準化番号に基づいて、前記正準化データを作成する第 3 のステップとを備える；

正準化データ作成方法。

4. 前記第 1 のステップは、

各原子に 3 種類の属性 (a_i , b_{ij} , d_{ij}) を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記 3 種類の属性 (a_i , b_{ij} , d_{ij}) の中で、 a_i は入力番号 i の原子の種類記号であり、 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が j である結合の数であり、 d_{ij} は入力番号 i の原子から最短経路により、 j 個の結合を経て巡れる道筋の数であり；

前記第 2 のステップは、

正準化番号を 1 から昇順に各原子に与える過程において、前記クラス番号の優先順位が最高である原子に正準化番号 1 を与え、以降正準化番号 n までが付与されているとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、前記クラス番号の優先順位が最高である原子に正準化番号 $n+1$ を与えており；

前記第 3 のステップは、

各原子に 3 種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一列に並べることによって前記正準化データを作成しており、

前記 3 種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号である；

請求項 3 に記載の正準化データ作成方法。

5. 化合物を構成する各原子についての固有データ及び原子間の結合対データの入力を受け付ける入力手段と、コンピュータ使用可能な媒体から情報を読み出す読出手段とを備える情報処理装置と共に用いるための正準化データ作成用コ

ンピュータプログラム製品であって；

プログラムを記録するためのプログラム領域を有しており、かつ、前記固有データ及び結合対データに基づいて正準化データを効率的に作成させるための該媒体中に具現化されたコンピュータ読取可能なプログラムを有しているコンピュータ使用可能な媒体を前記コンピュータプログラム製品は備えており；

該コンピュータプログラム製品は、前記プログラム領域において、前記固有データ及び結合対データに基づいて、前記化合物の化学構造を一意的に特定できる正準化データを作成する、コンピュータ読取可能な正準化データ作成プログラムを有しており；

前記正準化データ作成プログラムは、

前記固有データ及び結合対データに基づいて、各原子を等価原子ごとに別のクラスに分類して、クラスごとに異なるクラス番号を各原子に与える、コンピュータ読取可能な第1の処理ルーチンと、

前記第1の処理ルーチンで各原子に与えられたクラス番号に基づいて、前記化合物の構造と一意的に対応した正準化番号を各原子に与える、コンピュータ読取可能な第2の処理ルーチンと、

前記第2の処理ルーチンで各原子に与えられた正準化番号に基づいて、前記正準化データを作成する、コンピュータ読取可能な第3の処理ルーチンとを備える；

正準化データ作成用コンピュータプログラム製品。

6. 前記第1の処理ルーチンは、

各原子に3種類の属性 (a_i , b_{ij} , d_{ij}) を与えて、これらの属性が一つでも異なる原子は非等価であると判定できることを利用して、各原子を等価原子ごとに異なるクラス番号を付与しており、

前記3種類の属性 (a_i , b_{ij} , d_{ij}) の中で、 a_i は入力番号 i の原子の種類記号であり、 b_{ij} は入力番号 i の原子に隣接する結合のうち、その種類記号が

j である結合の数であり、 d_{ij} は入力番号 i の原子から最短経路により、 j 個の結合を経て巡れる道筋の数であり；

前記第 2 の処理ルーチンは、

正準化番号を 1 から昇順に各原子に与える過程において、前記クラス番号の優先順位が最高である原子に正準化番号 1 を与え、以降正準化番号 n までが付与されているとき、既に正準化番号が与えられている原子で、且つまだ正準化番号が与えられていない原子が結合している原子の中で、正準化番号が最小である原子を選び、その原子に結合している原子で、且つまだ正準化番号が与えられていない原子の中で、前記クラス番号の優先順位が最高である原子に正準化番号 $n + 1$ を与えており；

前記第 3 の処理ルーチンは、

各原子に 3 種類の属性 (P_i , T_i , S_i) を与えて、これらの属性を一行に並べることによって前記正準化データを作成しており、

前記 3 種類の属性 (P_i , T_i , S_i) の中で、 P_i は正準化番号 i の原子に結合し且つ正準化番号が最小の原子の正準化番号であり、 T_i は正準化番号 i の原子と正準化番号 P_i の原子との結合の種類記号であり、 S_i は正準化番号 i の原子の種類記号である；

請求項 5 に記載の正準化データ作成用コンピュータプログラム製品。

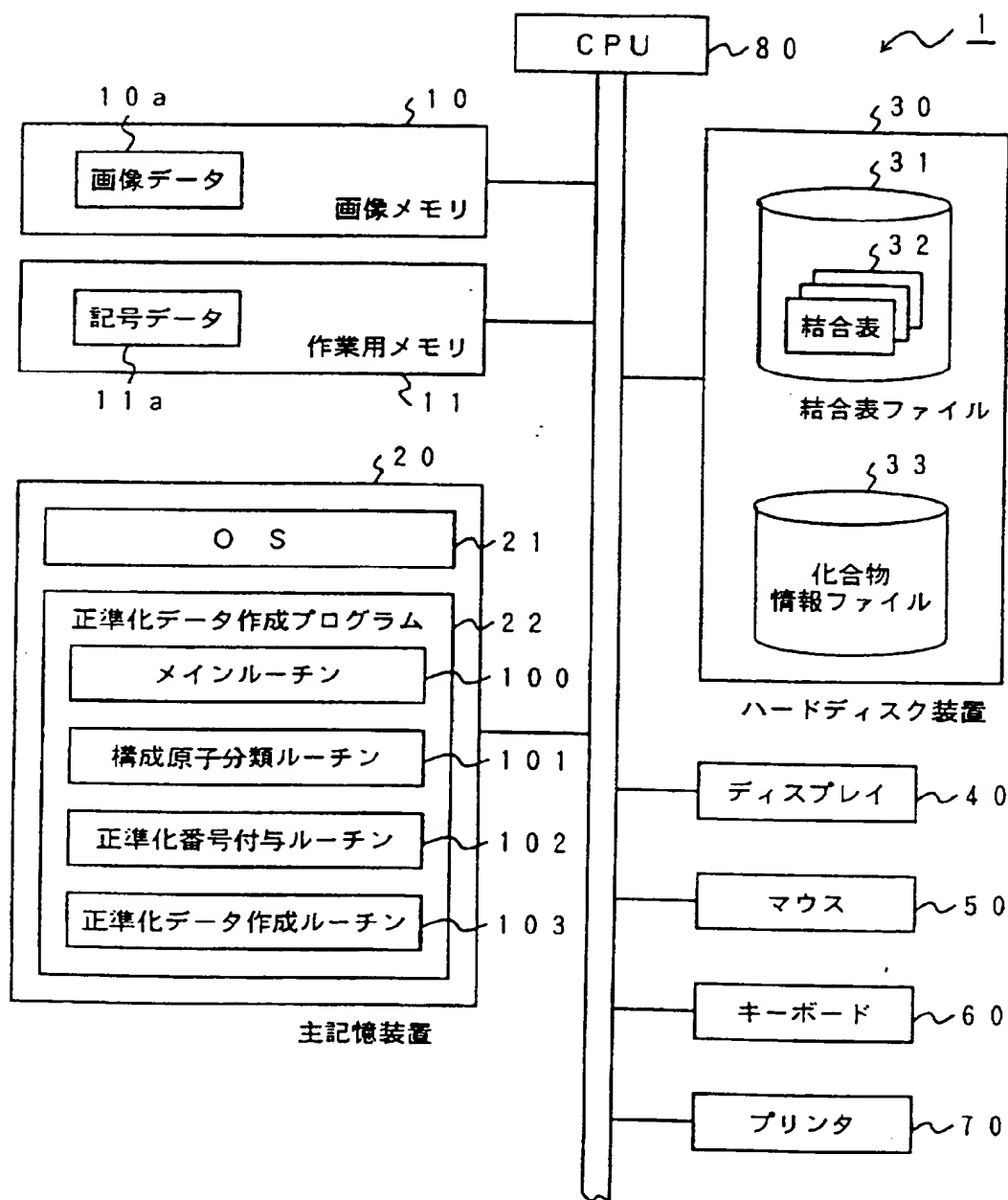


図 1

図 2 A

原子テーブル

3 2 a

入力番号	X座標	Y座標	元素名	属 性	原子数
1	3.0386	-2.4082	C		7
2	3.0386	-0.9082	C		
3	1.7396	-0.1582	C		結合数
4	0.4405	-0.9082	N		7
5	0.4405	-2.4082	C		
6	1.7396	-3.1582	C		
7	4.3376	-3.1582	C		

図 2 B

原子対テーブル

3 2 b

結合原子対		結合種	構 造
1	2	2	環
2	3	1	環
3	4	2	環
4	5	1	環
5	6	2	環
6	1	1	環
1	7	1	鎖

化合物情報ファイル

3 3

化合物番号	標準化データ	参照データ
C ₁	1 % 2 %	名称、文献、物性等
C ₂	
C ₃	
C ₄	
C ₅	
C ₆	
C ₇	

図 3

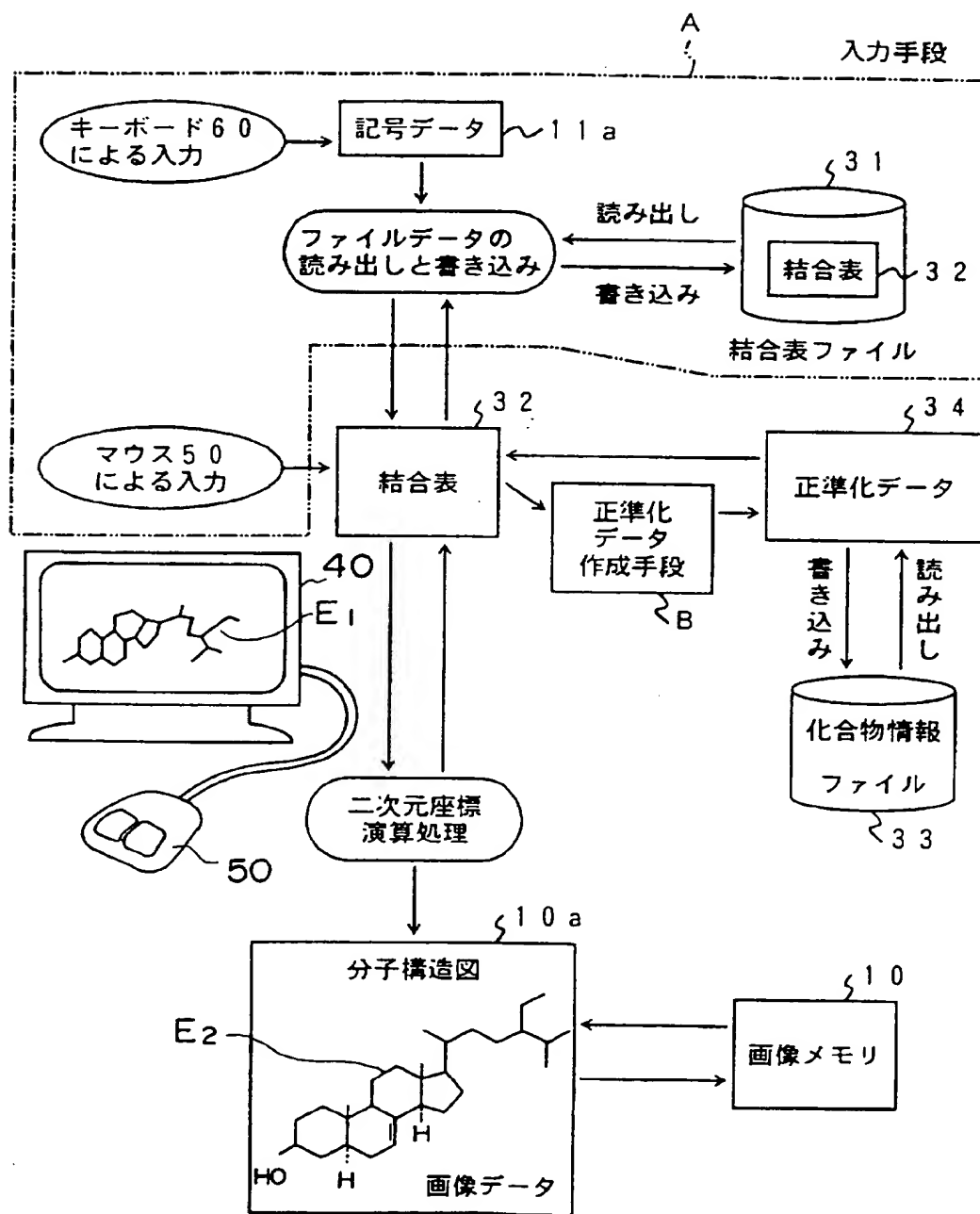


図 4

メインルーチン 100

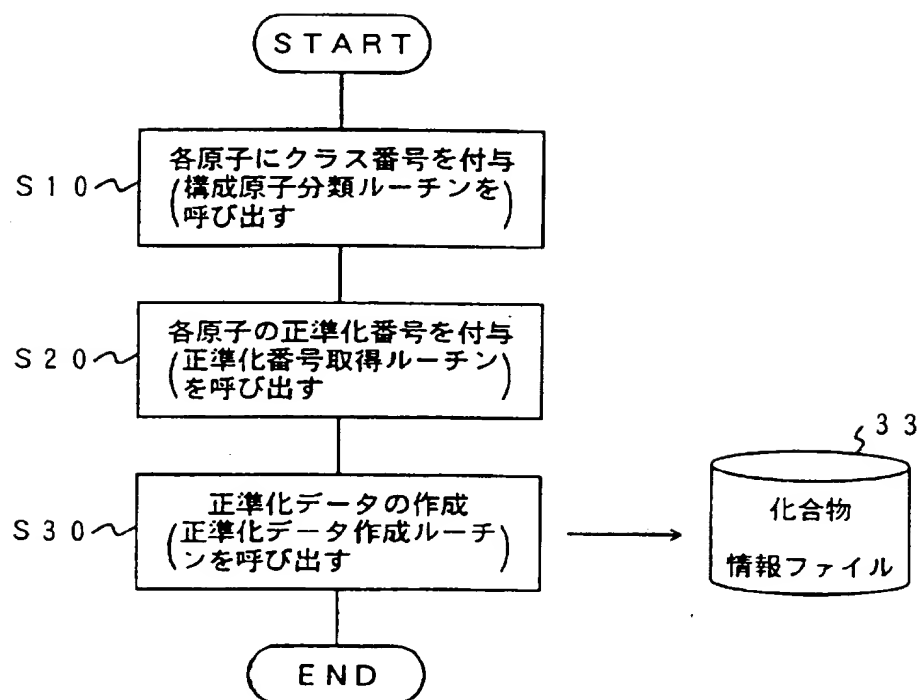


図 5

構成原子分類ルーチン 101

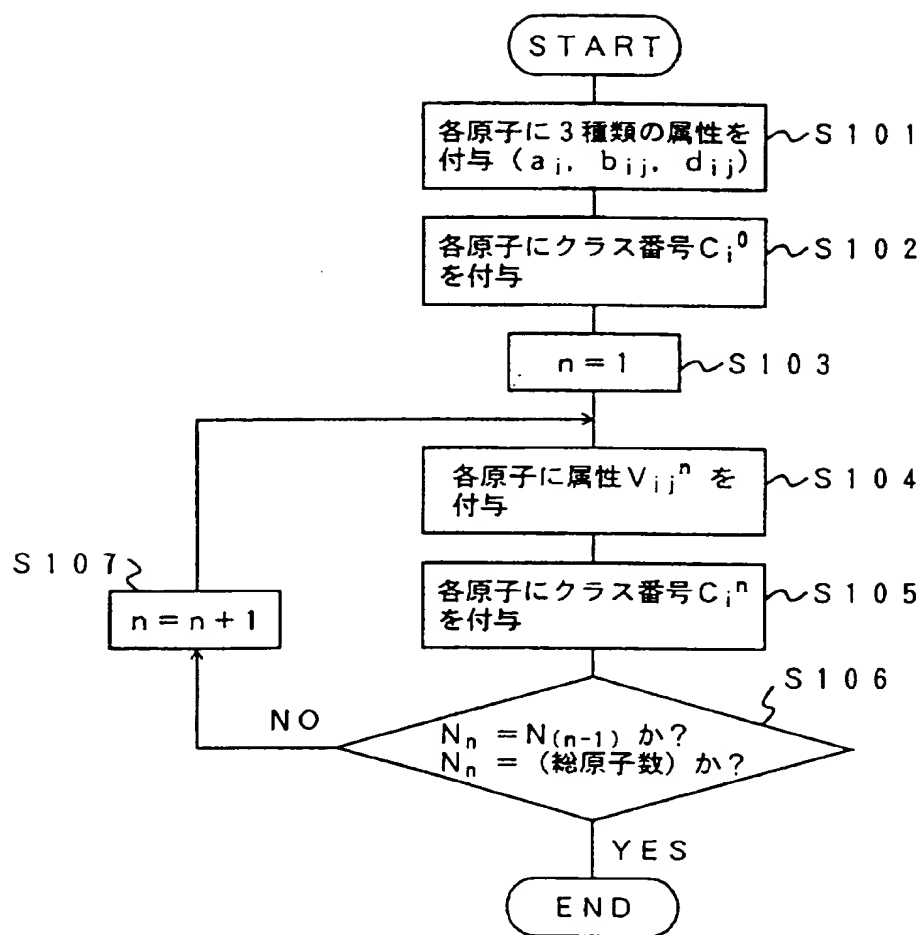


図 6

図 7 A

原子テーブル

3 2 a

入力番号	X座標	Y座標	元素名	属 性	原子数
1	—	—	C		8
2	—	—	C		結合数
3	—	—	N		
4	—	—	C		8
5	—	—	C		
6	—	—	C		
7	—	—	C		
8	—	—	C		

図 7 B

原子対テーブル

3 2 b

結合原子対		結合種	構 造
1	2	1	
2	3	2	
3	4	1	
4	5	1	
5	6	1	
6	1	1	
5	7	1	
1	8	1	

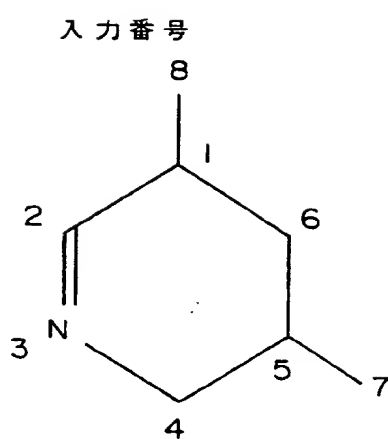


図 8

図 9A

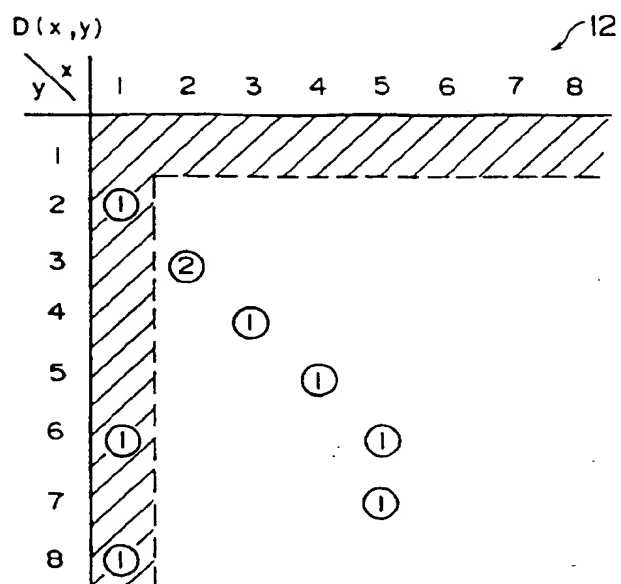
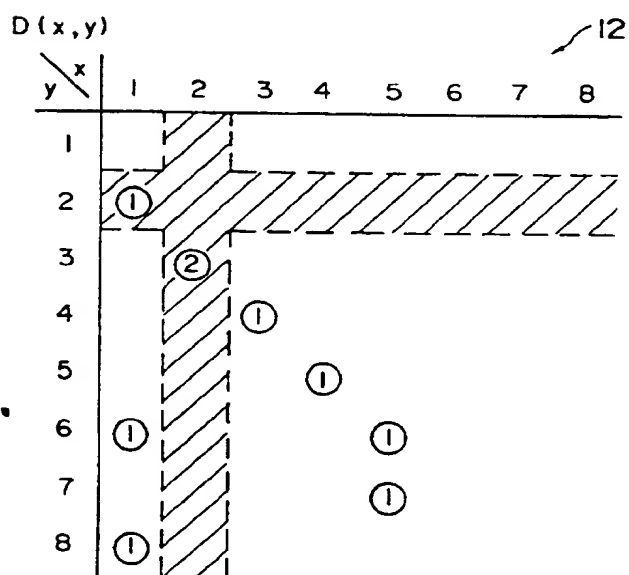


図 9B



$n = 0$

i	a_i	b_{ij}	d_{ij}	C_i^0
1	6	(3, 0, 0, 0)	(3, 2, 3, 0)	5
2	6	(1, 1, 0, 0)	(2, 3, 2, 2)	2
3	7	(1, 1, 0, 0)	(2, 2, 4, 0)	6
4	6	(2, 0, 0, 0)	(2, 3, 2, 2)	3
5	6	(3, 0, 0, 0)	(3, 2, 3, 0)	5
6	6	(2, 0, 0, 0)	(2, 4, 2, 0)	4
7	6	(1, 0, 0, 0)	(1, 2, 2, 3)	1
8	6	(1, 0, 0, 0)	(1, 2, 2, 3)	1

 $N_0 = 6$

図10

図11A

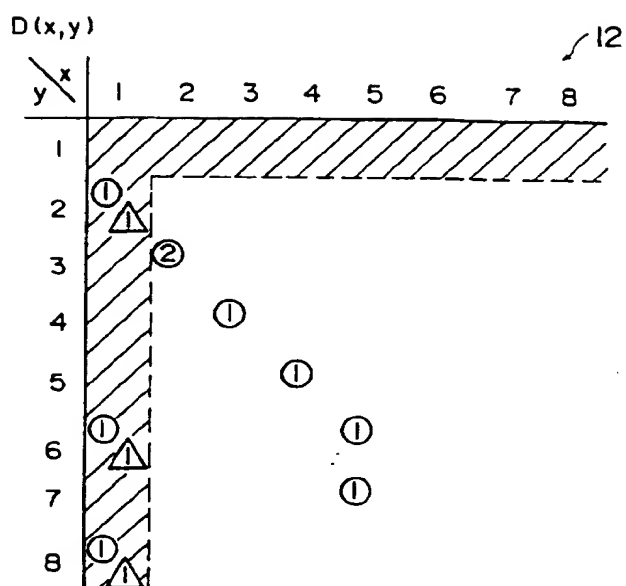
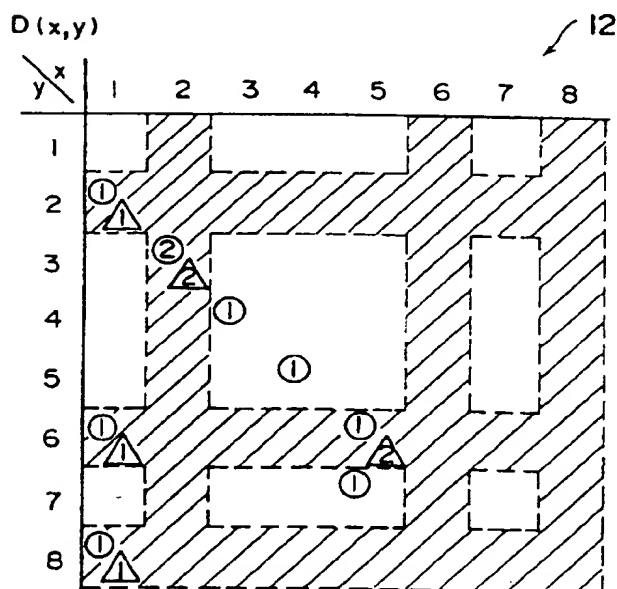


図11B



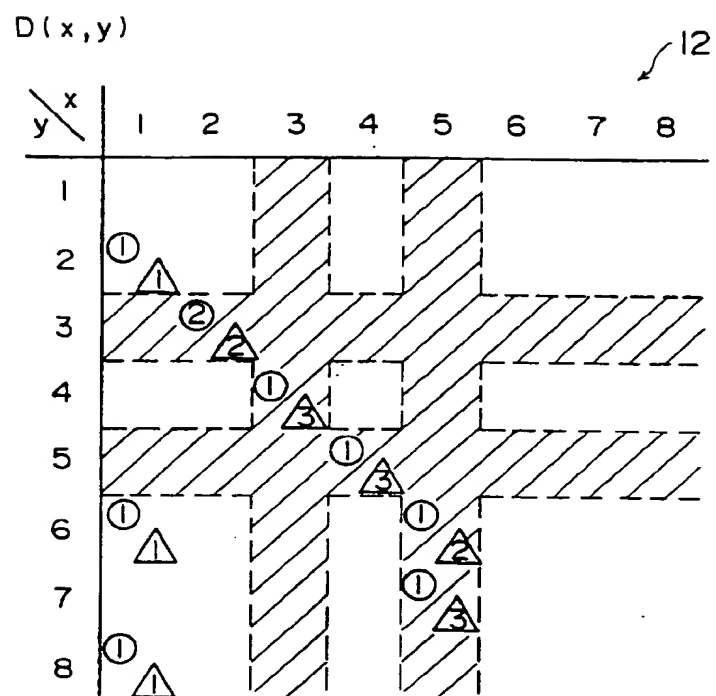


図12

図13A

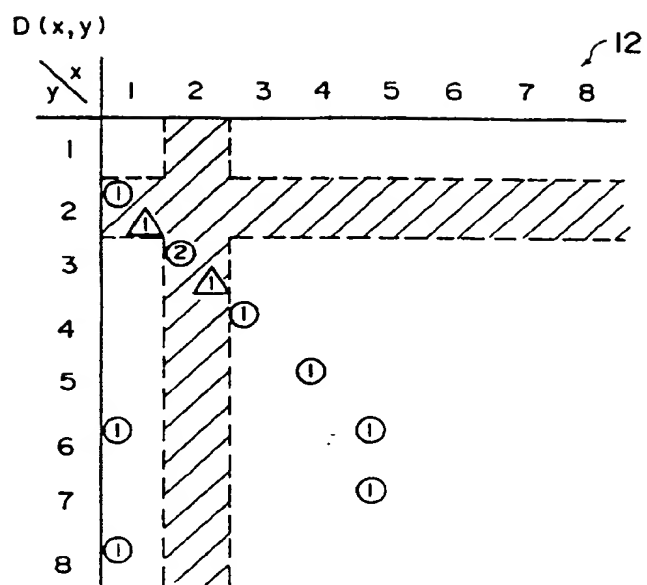


図13B

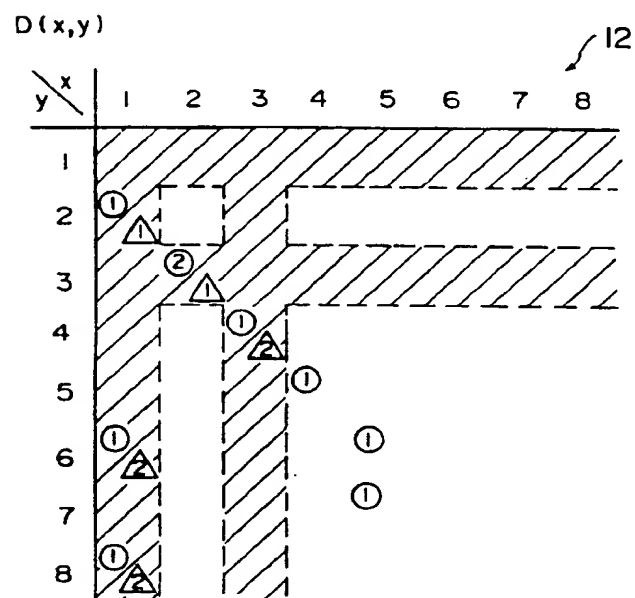


図14A

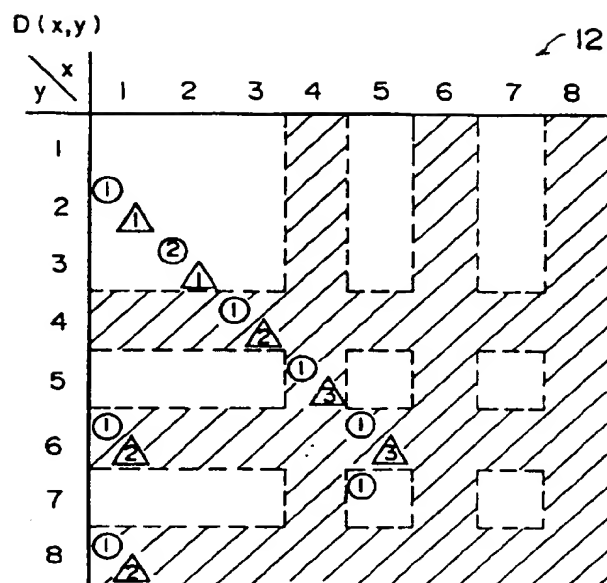


図14B

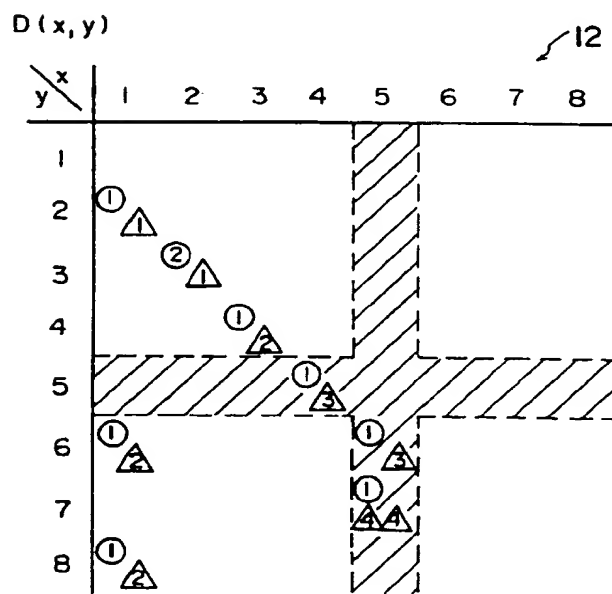


図15A

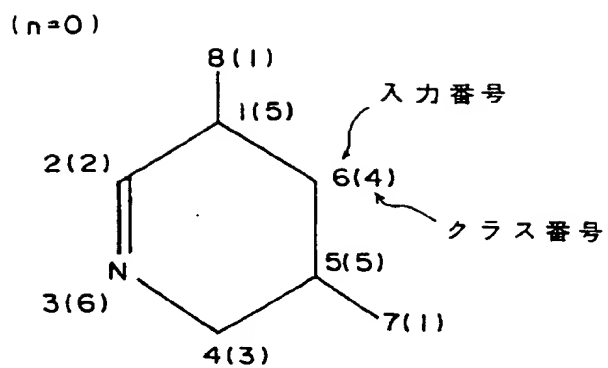


図15B

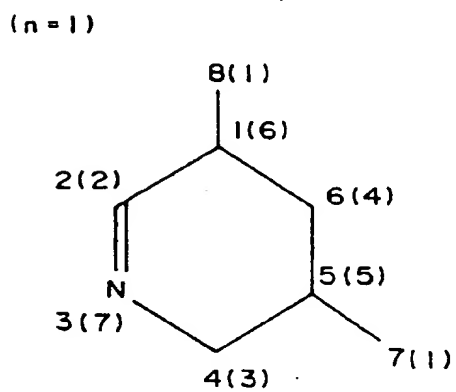
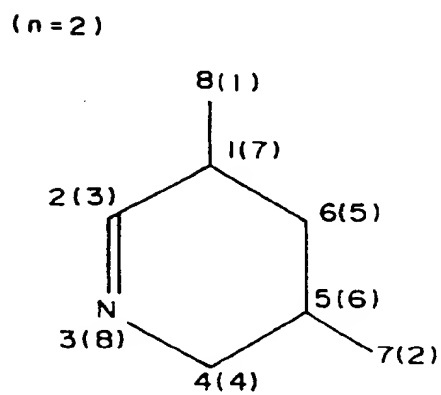


図15C



$n = 1$

i	V_{ij}^1	C_i^1
1	(1, 1, 0, 1, 0, 0)	6
2	(0, 0, 0, 0, 1, 1)	2
3	(0, 1, 1, 0, 0, 0)	7
4	(0, 0, 0, 0, 1, 1)	3
5	(1, 0, 1, 1, 0, 0)	5
6	(0, 0, 0, 0, 2, 0)	4
7	(0, 0, 0, 0, 1, 0)	1
8	(0, 0, 0, 0, 1, 0)	1

 $N_1 = 7$

図16

$n = 2$

i	V_{ij}^2	C_i^2
1	(1, 1, 0, 1, 0, 0, 0)	7
2	(0, 0, 0, 0, 0, 1, 1)	3
3	(0, 1, 1, 0, 0, 0, 0)	8
4	(0, 0, 0, 0, 1, 0, 1)	4
5	(1, 0, 1, 1, 0, 0, 0)	6
6	(0, 0, 0, 0, 1, 1, 0)	5
7	(0, 0, 0, 0, 1, 0, 0)	2
8	(0, 0, 0, 0, 0, 1, 0)	1

 $N_2 = 8$

図17

正準化番号付与ルーチン 102

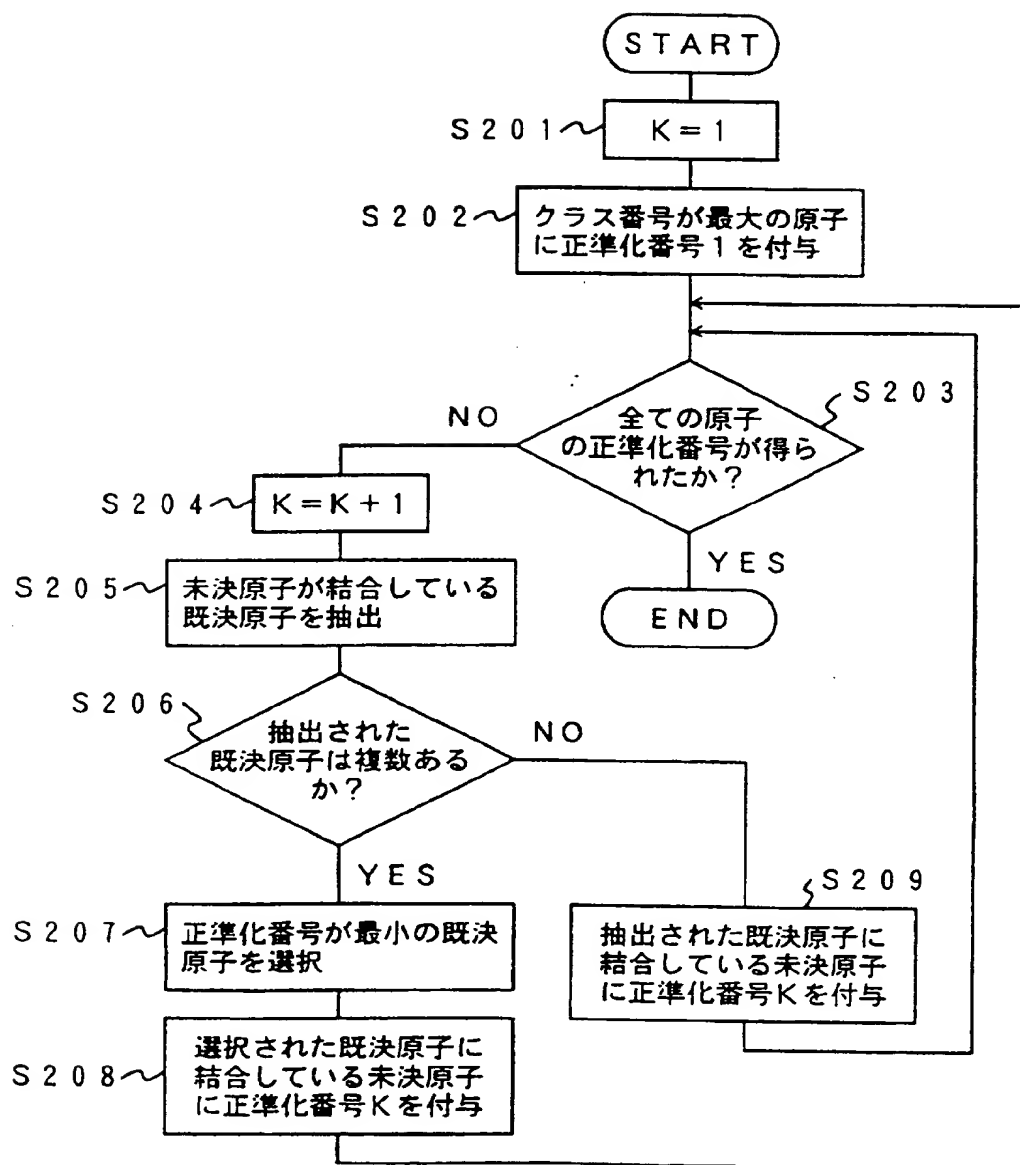


図18

正準化番号

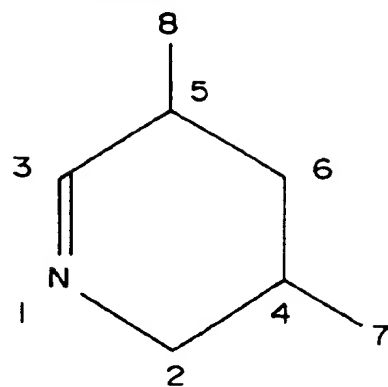


図19

正準化データ作成ルーチン 103

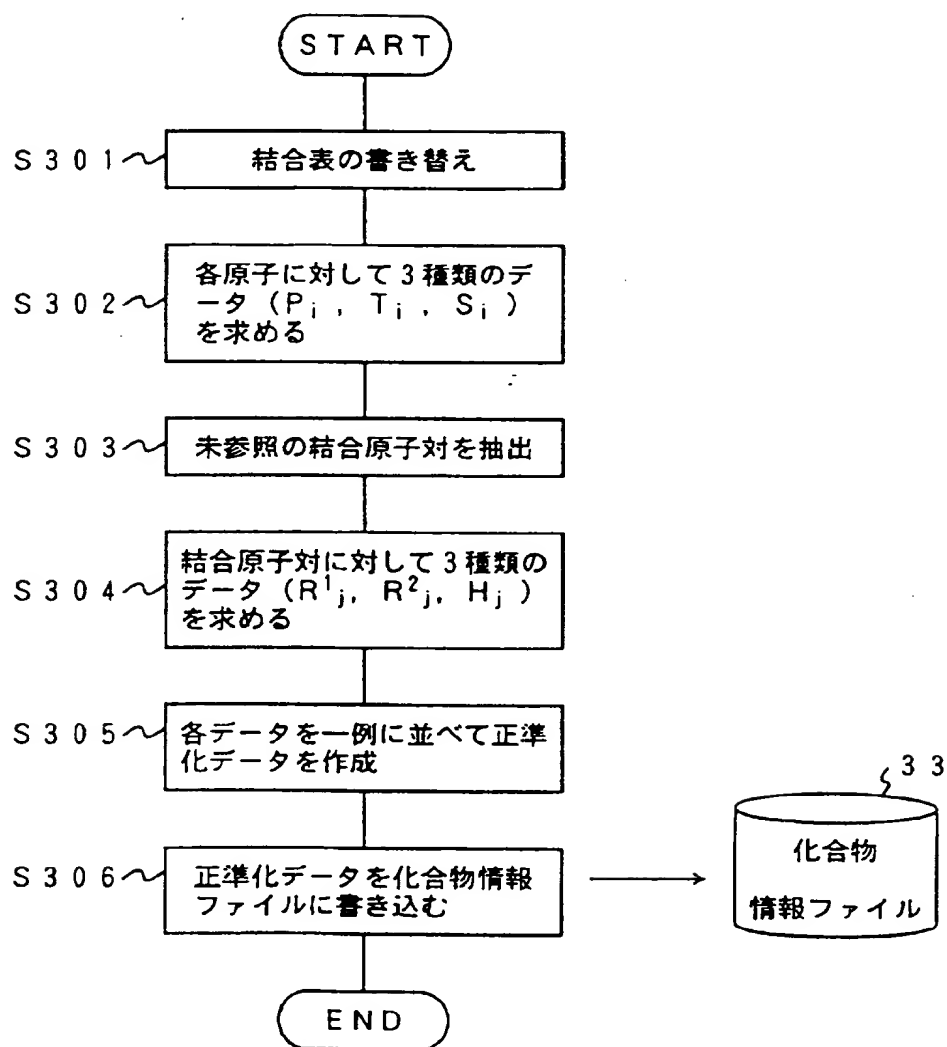


図20

原子テーブル

3 2 a

図21A

正準化番号	X座標	Y座標	元素名	属 性	原子数
1	—	—	N		8
2	—	—	C		結合数
3	—	—	C		
4	—	—	C		8
5	—	—	C		
6	—	—	C		
7	—	—	C		
8	—	—	C		

原子対テーブル

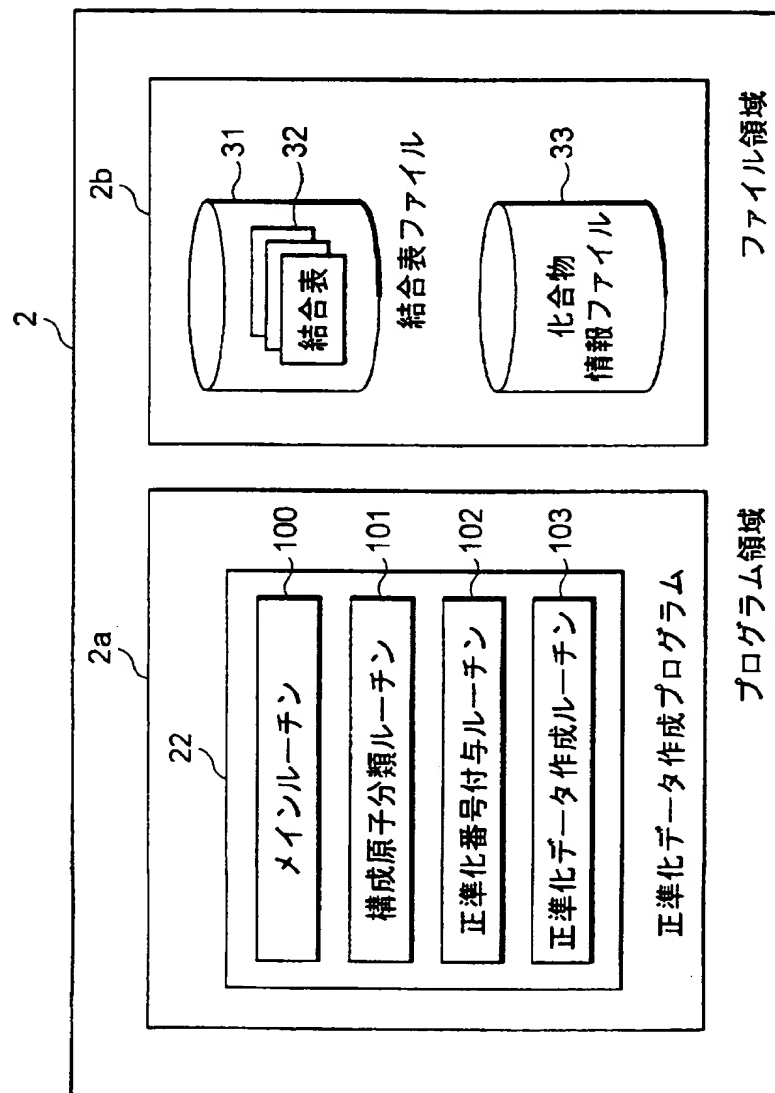
3 2 b

図21B

結合原子対		結合種	構 造
5	3	1	
3	1	2	
1	2	1	
2	4	1	
4	6	1	
6	5	1	
4	7	1	
5	8	1	

K		P_K	T_K	S_K
1		N
2		1	—	C
3		1	=	C
4		2	—	C
5		3	—	C
6		4	—	C
7		4	—	C
8		5	—	C
ℓ	R_ℓ^1	R_ℓ^2	H_ℓ	
1	5	6	—	

図22



正準化データ作成用記憶媒体

図23

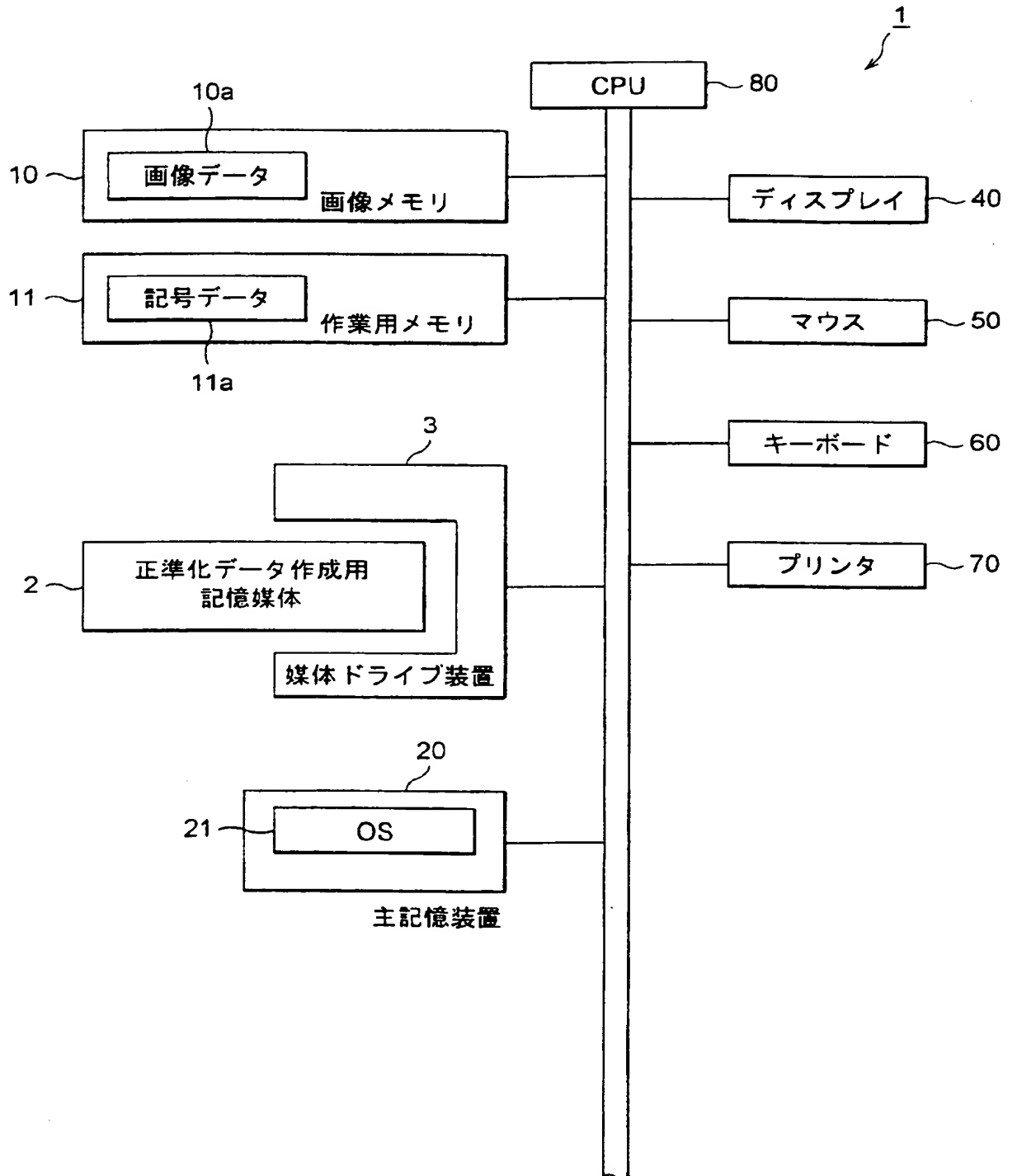


図24

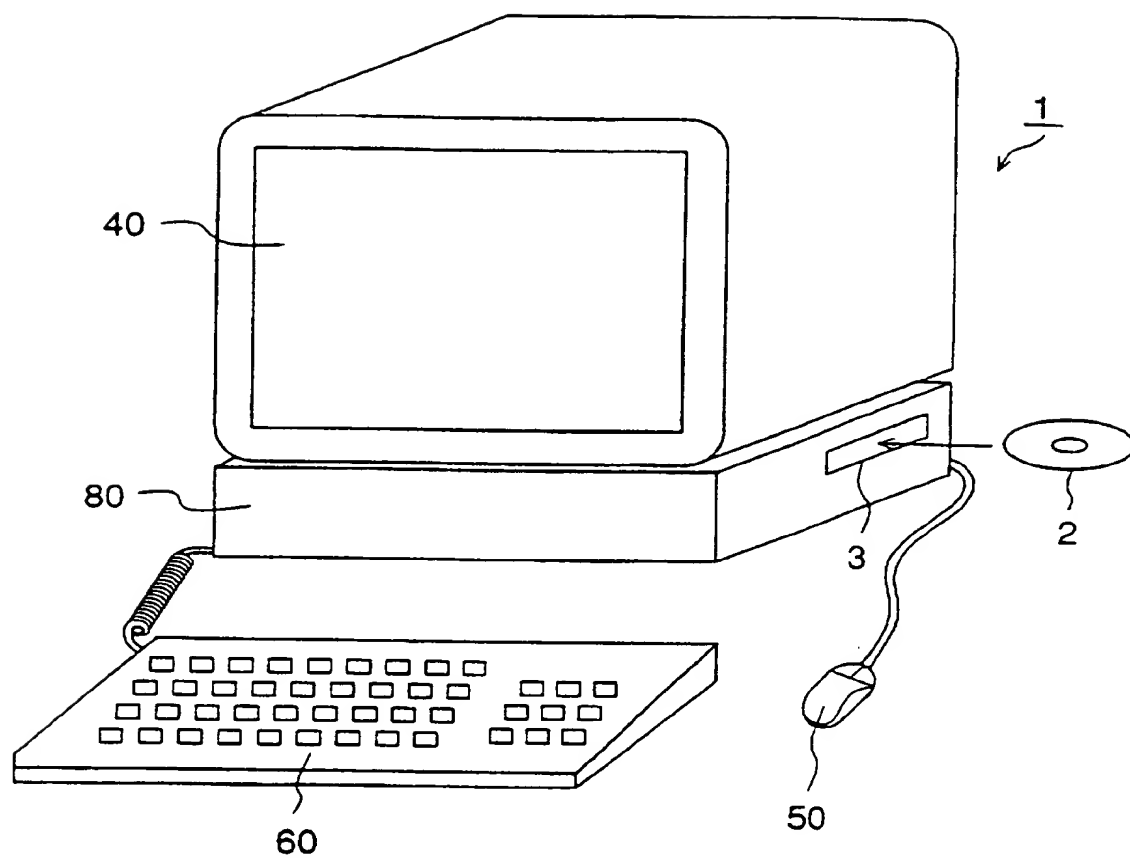
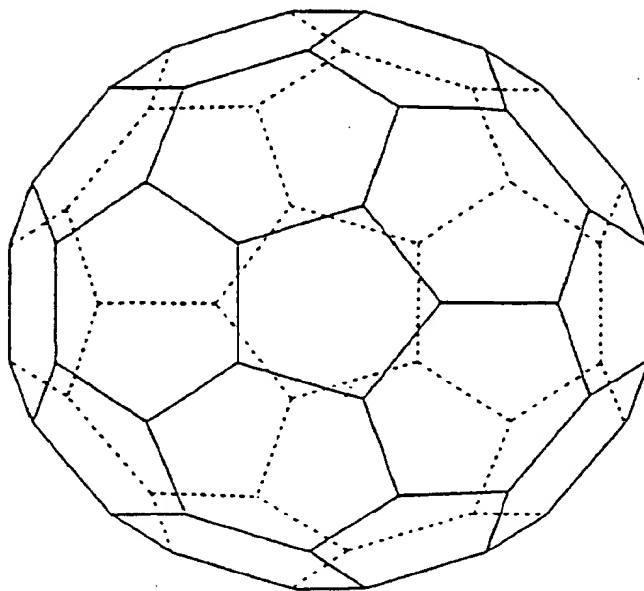


図25

画像データ 80e (C₆₀の分子構造図)

図26A



正準化データ 82d

図26B

1%1%1%2%2%3%3%4%4%5%5%6%6%7%8%9%10%1
 1%12%13%14%15%15%16%17%17%18%19%19%2
 1%21%23%24%25%26%27%28%29%30%31%32%3
 3%33%35%36%36%38%39%41%43%44%45%46%4
 7%48%49%50%53%57%/7%12/8%10/9%14/11%
 13/16%24/18%27/20%23/20%30/22%26/22%
 32/25%28/29%31/34%35/34%43/37%38/37%
 45/39%40/40%44/41%42/42%47/45%48/49%
 50/51%52/51%53/52%57/54%55/54%56/55%
 58/56%59/58%60/59%60/

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP97/01661

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl⁶ G06F17/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int. Cl⁶ G06F17/00, G06F17/40

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

JICST, WPI, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Akira Asanaga "Development and application of molecular structure processing programs and a library (in Japanese)" Symposium on Information Science Abstracts of the symposium of structure-activity relationship, Vol. 13th-18th, November 19, 1990 (19. 11. 90), p. 25-28	1, 2
A	WO, 96/06391, A (PSI International Inc.), February 29, 1996 (29. 02. 96), Page 9, line 1 to page 21, line 14 & US, 5577239, A & EP, 777882, A2	1, 2
A	JP, 4-98464, A (Fujitsu Ltd.), March 31, 1992 (31. 03. 92) & US, 5321804, A	1, 2
A	JP, 62-57017, A (Fuji Photo Film Co., Ltd.), March 12, 1987 (12. 03. 87) & US, 5056035, A	1, 2

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

August 6, 1997 (06. 08. 97)

Date of mailing of the international search report

August 19, 1997 (19. 08. 97)

Name and mailing address of the ISA/

Japanese Patent Office

Facsimile No.

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP97/01661

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 3, 4, 5, 6
because they relate to subject matter not required to be searched by this Authority, namely:
The claims 3 and 4 are related to the mental activity of a human being and the claims 5 and 6 are related to a computer program.
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int. Cl⁶ G06F17/00

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int. Cl⁶
G06F17/00, G06F17/40

最小限資料以外の資料で調査を行った分野に含まれるもの

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

JICST, WPI, BIOSIS

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
X	朝永惇「分子構造処理プログラム・ライブラリーの開発と応用について」情報化学討論会・構造活性相関シンポジウム講演要旨集, Vol. 13th-18th, 19. 11月. 1990 (19. 11. 90) p25-28	1, 2
A	WO, 96/06391, A (PSI International Inc.) 29. 2月. 1996 (29. 02. 96), 第9頁第1行-第21頁第14行&US, 5577239, A&EP, 777882, A2	1, 2
A	JP, 4-98464, A (富士通株式会社), 31. 3月. 1992 (31. 03. 92) &US, 5321804, A	1, 2
A	JP, 62-57017, A (富士写真フイルム株式会社), 12. 3月. 1987 (12. 03. 87) &US, 5056035, A	1, 2

☐ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」先行文献ではあるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日

06. 08. 97

国際調査報告の発送日

19.08.97

国際調査機関の名称及びあて先

日本国特許庁 (ISA/JP)

郵便番号100

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

相田 義明

5L

7925

電話番号 03-3581-1101 内線 3562

第I欄 請求の範囲の一部の調査ができないときの意見 (第1ページの1の続き)

法第8条第3項 (PCT 17条(2)(a)) の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1. ☒ 請求の範囲 3, 4, 5, 6 は、この国際調査機関が調査をすることを要しない対象に係るものである。つまり、

請求の範囲 3, 4 は、人間の精神活動に関するものであり、請求の範囲 5, 6 はコンピュータプログラムに関するものである。

2. ☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、

3. ☐ 請求の範囲 _____ は、従属請求の範囲であってPCT規則6.4(a)の第2文及び第3文の規定に従って記載されていない。

第II欄 発明の単一性が欠如しているときの意見 (第1ページの2の続き)

次に述べるようにこの国際出願に二以上の発明があるとの国際調査機関は認めた。

1. ☐ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。
2. ☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。
3. ☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。
4. ☐ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意

- ☐ 追加調査手数料の納付と共に出願人から異議申立てがあった。
- ☐ 追加調査手数料の納付と共に出願人から異議申立てがなかった。